Design and Evaluation of Diagnostic Studies

Werner Vach, Veronika Reiser, Izabela Kolankowska, Susanne Weber, Gerta Rücker

April 24, 2017

Contents

Preface

Preparing a first course about the design of diagnostic studies in the year 2014, and also revising it in 2017 has been a challenging task. Our view on diagnostic studies has changed during the last years. Guideline developers and health policy makers have started to require the same standards of evidence based medicine in diagnostic research as in therapeutic research. In particular, this has questioned the role of accuracy studies as the final step in developing new diagnostic tests and has put the focus on studies or other approaches to provide information on the patient benefit implied by using a diagnostic test. Schünemann et al. (2008a) put it simply into the words 'If a test fails to improve patient-important outcomes, there is no reason to use it, whatever its accuracy.'

Thus the field of diagnostic studies is in a transition, putting more and more emphasis on benefit on top of accuracy, but also sharpening the general quality standards for diagnostic studies. This implies that many issues in this transition are at the moment a matter of debate. Any participant of this course should have this in mind when going through the material, and they should always remember that there are many points where they have to build up their own opinion. There are of course some basic facts which are not debatable, but you will find that we often can only argue that some ideas are more convincing than others. And then you have to make up your mind whether you can agree on this or not.

As with any course script, we are faced with the problem of introducing a complex topic, which is sometimes hard to be pressed into a linear order. We tried to solve this problem by allowing at some places repetitions of what has already been said, and some forward links to later chapters, which may clarify a point left open before. Each chapter starts by stating its objectives and ends with a summary. At the end of some chapters we added a few remarks, typically including information possibly of interest to the reader, but not obligatory for the further course. At some places we added hints to further reading. They mainly refer to articles or books providing a more in depth discussion of the points raised or presenting an alternative view. It is not expected that you read these articles as part of the course. They are only meant for participants who would like to deepen their understanding beyond the level of this course.

Freiburg, April 24, 2017

Werner Vach
Veronika Reiser
Izabela Kolankowska
Susanne Weber
Gerta Rücker

Part A

The Basics

Chapter 1

What is a Diagnostic Test, and what should we Know about it?

Objectives of Chapter 1

At the end of Chapter 1 the reader should be able to ...

- describe the general structure of a diagnostic test
- briefly describe the accuracy of a diagnostic test
- define the so-called gold standard
- understand the role of the gold standard for determining the accuracy of a diagnostic test
- know sensitivity and specificity as measures for accuracy
- distinguish between accuracy and benefit
- recognize that randomized controlled trials (RCTs) are a tool to assess the benefit of a diagnostic test

The basic common characteristic of a diagnostic test is the structure of the results: It is a dichotomous decision (or better suggestion) in favor or against a specific disease state. The type of the two disease states to be distinguished can range from the global distinction 'diseased' vs. 'disease-free' to very subtle characteristic of a disease, for example the presence or absence of a mutation. Table 1.1 summarizes typical examples. Also the type of the diagnostic test

diseased	disease-free
severe disease	mild disease
stage II	stage I
lymph nodes affected	lymph nodes unaffected
receptor positive	receptor negative
tumor size $>$ 5cm	tumor size $<$ 5cm
receptor positive	receptor negative
tumor nonresectable	tumor resectable
mutation present	mutation absent

Table 1.1: Examples of disease states to be distinguished by using a diagnostic test

can range from very simple procedures like a single symptom or a single question ('Is your pain located at the left side or the right side of the abdomen?') to very complicated algorithms summarizing the results from many single tests or measurements. Table 1.2 summarizes typical examples of types of diagnostic tests. Often, diagnostic tests consist of two components: Some type of (technical) measurement procedure to generate information and some type of decision or decision rule to obtain a dichotomous result, which may include subjective elements like the interpretation of an image. We will later see that the internal structure of a test may have some impact on the planning of diagnostic studies, but for many considerations the internal structure does not matter. Hence in the moment it is sufficient to focus on the general structure of a diagnostic test, namely that it makes a suggestion for distinguishing between two disease states.

The first basic question we can ask about a diagnostic test is: How good is the test in distinguishing between the two disease states? This is called the accuracy of the test. It can be studied if we have another test allowing us to determine the true disease state of each patient. Such a test is called a gold standard. Hence the basic idea of any accuracy study is to apply both the test of interest and the gold standard in a series of subjects and to study empirically the accuracy. Typically, this is approached by computing two numbers, namely the percentages of correct test decisions by the test of interest separately for the two disease states of interest as identified by the gold standard. In the simple case of distinguishing between diseased and disease-free subjects, these two percentages are known as sensitivity and specificity:

Type of test	Example
single symptom	abdominal pain
single question	'Is your abdominal pain located at the right body side?'
clinical measurement with cut-off	body temperature > 38,5
laboratory parameter with cut off	cardiac troponin T > 0.01 ng/mL
image with visual interpretation	x-ray and detection of pneumonia
image with extraction of parameter	SUV(max) values based on PET/CT images
image with semiquantative parameter	Summary stress score in myocardial perfusion imaging (SPECT): An image of the heart is divided into 20 segments and each segment is graded on a scale from 0 to 3.
symptom scale with cut off	PHQ-9: depression subscale of the Patient Health Questionnaire. Cut point for severe depression: 20
gene expression measurements with algorithm	Oncotype DX (trademark): Prediction of response to adjuvant chemotherapy in breast cancer patients

Table 1.2: Examples of types of diagnostic tests

Sensitivity is the percentage of subjects classified as 'diseased' among those who are diseased, and specificity is the percentage of subjects classified as 'disease-free' among those who are disease-free. Accuracy studies play an important role in diagnostic research and account today for the vast majority of all diagnostic studies. Consequently, we will discuss their design in depth in this course.

However, seen from a broader perspective, the concept of accuracy lacks one important aspect: the benefit for patients. An improved accuracy alone does, indeed, not imply any benefit for patients. The benefit arises from a change in treatment decisions, or – more generally speaking – a change in the management of the patients. Only if improved accuracy results in improved management of patients we can expect that improved accuracy results into a benefit for patients. And there are – as we will discuss in more detail in 3.1 – sometimes good reasons to have some doubts about this translation.

Today, benefit is a key concept for guideline developers as well as for health technology assessment (HTA) agencies informing decision makers who have to decide on introducing new diagnostic modalities in routine care or the corresponding reimbursement. Hence guideline developers and HTA agencies require today often evidence for such a benefit for patients, and accuracy studies alone do not provide such an evidence base.

This has brought another type of diagnostic studies into the focus, namely studies which try to assess directly the benefit for patients. In particular, randomized controlled trials (RCTs) directly comparing two different diagnostic tests using patient centered outcomes have been advocated, as they allow to establish an evidence base similar as in therapeutic research, where RCTs often play a central role. In its most stringent form, randomized diagnostic studies try to answer a question like: Does the use of diagnostic test A instead of diagnostic test B improve the overall survival of patients? It is not surprising that such a radical change in study culture from (small, single-center) accuracy studies to (large, multi-center) RCTs raises a lot of questions and confusions. It will be one of the aims of this course to give a structured introduction to the topic of benefit studies and in particular to present an overview about the many designs proposed so far and their advantages and limitations. Hence besides accuracy studies benefit studies will be the second major topic of this course. Further study types will be only briefly mentioned in 4.

The broader perspective of benefit also allows us to investigate diagnostic approaches which go beyond the concept of a single diagnostic test. For example, a single image is today often not only used to make one single decision, but to inform a whole treatment like radiation therapy or a complicated surgery. It may be even used to inform a patient about his or her disease and may serve as a basis for the communication with the patient. Accuracy is not a concept which can measure the impact of all these consequences of a single image, but the

concept of benefit allows this. For example we can still compare the benefit of two diagnostic modalities (e.g., PET/CT and MRI) in a randomized trial, even if they are allowed to be used several times to support various decisions and communication processes. Other examples arise when diagnostic processes are interacting with patient preferences. Whenever we consider such general approaches which go beyond a single diagnostic test, we will use the term 'diagnostic procedure' instead of 'diagnostic test'. It is rather obvious that if we talk about the patient benefit of a diagnostic test or a diagnostic procedure, we are actually talking about the benefit of a specific combination of diagnostic tests or procedures with a choice from certain therapy options. This strong connection between diagnosing and treatment is, however, nothing new. This strong connection has always been the reason for requiring diagnostic tests to result in only two possible states 'positive' and 'negative', although in clinical practice a middle category like 'unsure' may be highly adequate. However, as diagnostic tests should support treatment decisions (and as long as there are only two treatment options), they have to give a definite, dichotomous answer. Otherwise, they are not clinically useful.

Summary of Chapter 1

A diagnostic test provides a dichotomous decision rule for distinguishing between two disease states. How good a specific test can distinguish between the disease states of interest is investigated in accuracy studies. The key element in accuracy studies is the application of a gold standard test determining the true disease state of each patient. Beyond diagnostic accuracy studies, randomized diagnostic studies are appropriate. They concentrate on the potential benefit patients may have from a change in treatment decisions or management by/after applying the test.

Chapter 2

Basic Issues in Designing Accuracy Studies

Objectives of Chapter 2

At the end of chapter 2 the reader should be able to ...

- understand the aim of a diagnostic accuracy study
- describe what is meant by target condition
- name at least one more common term for the gold standard
- describe what is meant by index test
- differentiate between reference test and index test
- specify potential impacts to the generalizability of accuracy studies
- describe what is meant by target situation
- differentiate between target population and study population
- have a first idea about definitions and concepts of different measures of accuracy

As mentioned above, the basic idea of any accuracy study is quite simple: We apply the test of interest and the gold standard in a series of patients, and we observe empirically the accuracy of the test. However, if we want to perform an informative and convincing accuracy study more considerations are necessary. As with many other medical studies the basic aim is to inform clinicians, patients and other stakeholders about what they can expect in daily clinical routine if a certain intervention, procedure or test is used. This requires that the results of the study are generalizable to clinical routine, or that it can be at least judged to which degree they are generalizable. Key questions to be addressed to allow such a judgment from an accuracy study are

- How was the test actually performed?
- In which population did we apply the test?
- What does the gold standard actually identify?
- How do the numbers computed to describe the accuracy relate to clinical practice?

So these are the question we will address in this chapter, as they are independent of the study design used to perform an accuracy study, and as they apply to a large degree also to benefit studies. The question of the actual choice of the study design for an accuracy study will be addressed in Chapters 5 and 10.

2.1 Target Condition, Gold Standard and Reference Test

Among the two disease states we would like to distinguish, there is often one which reflects our main interest. This state is often called the 'target condition', for example the presence of a disease or the presence of metastases. Often, the definition of the target condition seems to be rather clear, as it reflects the aim of the diagnostic test. However, we do not only need a clear idea of how to define the presence of the target condition – under the assumption that we know everything about the true status of the patient – but also how to define its absence. Actually, we have often some patients which are somewhere between the absence and the presence of the target condition. For example, should the target condition 'myocardial infarction' include 'silent' myocardial infarctions with no chest pain, or should it include coronary thrombosis reversed by thrombolytic treatment, which then averts full infarction?

If we have a disease (or disease state) with a spectrum ranging from mild to severe, then in any case the definition of the target condition requires to choose a cut point. Such a cut

point may be a single numerical value as in the case of obesity (Body Mass Index (BMI)>30) or hypertension (blood pressure > 140/90mmHg). In other areas, such a cut point may be the presence of certain key signs. It might be also necessary to exclude explicitly related diseases or disease states, i.e., to require the absence of certain signs.

In any case, at the end we need a clear conceptual definition for the presence and absence of the target condition, allowing us to classify all patients if we would know everything about their true status.

Next we need a gold standard, i.e., a test allowing us to determine the presence or absence of the target condition in each patient. At first sight the existence of such a gold standard seems to be counter intuitive: Why should we evaluate a new diagnostic test if we have already a perfect test? However, many gold standards arise from the simple fact that we often know the true disease state only later, when it is too late. For example, subtypes of Parkinson's disease can be easily determined by inspection of the brain after the patient died. The 1979 WHO criteria of myocardial infarction rely mainly on the results of a 24h ECG (Anonymous 1979), but treatment has to be initiated as soon as the patient presents him- or herself with chest pain. Another source of gold standards is given by invasive procedures, which cannot be justified in clinical routine, but which may be justified in a research context. For example findings from imaging procedures looking at anatomical structures can be in principle validated by some type of surgery, but it is exactly the aim of imaging procedures to avoid such surgeries for purely diagnostic purposes.

Although we have these theoretical arguments about the existence of gold standards, perfect gold standards are rather rare. Partially, this is a consequence of the difficulties we have mentioned already above with respect to defining the target condition: We have always some patients at the borderline between absence and presence. For example the 2000 WHO criteria for myocardial infarction require a 'typical rise and gradual fall (troponin) or more rapid rise and fall (CK-MB)' Alpert et al. (2000), such that we will always have patients where it can be debatable whether the rise is typical or more or less rapid. Another reason for lacking real gold standards arises from the fact that we often cannot handle all subjects equally once we know the result of the diagnostic test, as the result determines the further management of the patient. And this difference in management has an impact on the opportunities to validate the test results. As an extreme example, let us consider a diagnostic test to inform about whether a tumor is resectable or not, i.e., whether we can remove it by surgery or not. If the test is positive surgery is performed, and the surgeon can validate whether the tumor is resectable or not. If the test is negative, no surgery is performed, and we will never know the true status. Fortunately, in most situations the difference is less extreme. We have two different ways to confirm positive or negative test results with different degree of validity. For example, when

distinguishing between presence and absence of a disease, a positive test result implies often additional diagnostic tests allowing us to identify the incorrect positive test results, such that at the end the treatment decision is made in a state of clear evidence in favor for or against the presence of the disease. Or at the end treatment failure shows us that the disease was not present. In contrast, a negative test result may imply that the patient is sent home and the only source to confirm the negative test result is the follow up of the patient. If the patient comes back soon with (additional) symptoms, and we now come to the conclusion that the disease is present, we may regard the original negative test result as incorrect. If the patient never comes back and lives for many years without any signs of the disease, we regard the original negative test result as correct. However, the use of the follow up of a patient for confirmation is not error-free. If the patient develops (new) symptoms, we cannot be sure that the disease was already present at the day of the test. If the patient lives for many years, the disease may have been present at the time of diagnosis, but may have vanished without treatment, or a treatment was initiated, but we just do not know this.

Due to the absence of perfect gold standards it has become popular in the last years to use the term 'reference standard' or 'reference test'. The reference standard should reflect the best possible, clinical practice today, i.e., the optimum we can achieve today without artificial improvements which are ethically or economically not justifiable. Often, a reference standard is based on combining different sources of information, for example, follow up and additional lab tests in test negative patients and clinical verification in test positive patients. To allow the judgment of generalizability mentioned above, the reference test should be defined as detailed as possible and described accordingly in the publication.

Of course, using the phrase 'reference test' instead of 'gold standard' does not solve any problem implied by the imperfectness of this test. We will later in Section 12.1 discuss the consequences of an imperfect reference standard on the results of an accuracy study.

Remark: In many situations it is rather arbitrary which of the two disease states we would like to distinguish is selected as 'target condition'. Both disease states are typically of importance for the further management of patients. Selecting one state as the target condition is hence often just a tradition allowing us to continue to use terms like 'sensitivity' and 'specificity'. Consequently, often the more serious state is regarded as the target condition. This implies that the terminology of test results continues to lack the patient perspective: Most 'positive' test results are actually rather negative for the patient.

Remark: Besides the term 'gold standard', also the term 'golden standard' has been used in the literature. Sometimes this has been justified with arguments similar to why today the term 'reference test' is preferred.

2.2 The Index Test

In diagnostic research, the diagnostic test we actually want to evaluate is often called the index test to distinguish it from the reference test. To ensure the generalizability of an accuracy study the description of the index test must be sufficient to allow to reproduce it exactly in patients outside of the current study.

Many diagnostic tests include a technical part, for example the application of a certain assay or a certain imaging procedure. This technical part involves typically a certain instrument or assay, and we can refer to the exact name of the instrument/assay used and its manufacturer, who is then responsible for producing instruments with sufficient reproducibility of results. The technical part depends often on additional technical parameters we can typically standardize and describe rather precisely, for example the amount of blood used or the imaging time used. However, there is often still some room for variation, for example how the patient is prepared for the measurement (fasting prior to investigation, physical placement of patient etc.), at which time of the day the procedure was applied, or whether other tests were performed simultaneously. So these aspects have to be standardized, too. Similar consideration about standardization may apply to non-technical instruments like questionnaires, where the final accuracy may depend on how a patient is instructed to fill it in.

The most crucial point in many diagnostic tests is, however, the existence of subjective elements in the test. Images have to be interpreted to come to final test results, or symptoms have to be graded before they can enter a decision rule. Such subjective elements are always a thread to the generalizability of the results of an accuracy study. In particular if only one person is performing all tests in an accuracy study (and if this person has an outstanding, long experience with the test), we may raise the question whether the results relate in any way to clinical routine conditions.

There are several ways to reduce the potential impact of subjective elements on the generalizability of an accuracy study. First we can ensure that the tests are performed by several subjects with characteristics similar to typical test users in clinical routine. Examples of such characteristic can be medical specialization, education, or years of experience. This way we can try to mimic the situation we will later meet in practice with different testers. Second we can try to minimize the subjectivity by writing detailed manuals with a lot of examples, using them in the accuracy study as guideline for the testers, and making these manuals publicly available, such that they can also serve as a basis in clinical routine. Third, we can include some formal training of the testers as part of the accuracy study, such that the results refer to the accuracy we can obtain if this certain amount of training has been given. Of course, the training has to be described in detail, and in the best case a training manual is used and made publicly

available. A forth, but more problematic approach is to use several persons who independently interpret the results of each patient and then agree on a consensus test result. This can indeed increase the accuracy of a test, but it may introduce an artificial element, if later in clinical routine the results are only judged by one person.

2.3 Target Situation, Target Population, and Study Population

The accuracy of a test is not a constant. It depends on the composition of the population in which the diagnostic test is applied. For nearly all disease states we typically have subjects who are easy to diagnose and subjects who are difficult to diagnose. Subjects easy to diagnose are those for which the target condition is present and which have very clear and distinct symptoms, and those for which the target condition is not present and which shows absolutely no symptoms. Subjects difficult to diagnose are those for which the target condition is present, but showing very few or indistinct symptoms (for example because they are in a very early stage, or because they have suffered from the disease for many years and have been already adapted to it), and those for whom the target condition is absent, but who show symptoms (for example because they have a disease or disease state similar to but different from the target condition). Now there is a simple relation between the composition of a population with respect to the difficulty to come to a diagnosis and the accuracy of a test: The more subjects difficult to diagnose are in a patient population, the lower will be the accuracy of a test – as long as it is not a perfect test.

To illustrate this point, let us consider an allergy test for a common substance like hazelnuts. If such a test is applied in the population of a specialized allergy center of a university hospital, we should not be very surprised if the accuracy is only moderate. Probably, mainly subjects with inconclusive results from previous tests or other complications are sent to such a center, and hence in this patient population we have to expect mainly patients who are difficult to diagnose. In contrast, if a pediatrician working in a general care setting is applying the test in young children, he or she will probably experience a high accuracy, if many children are tested for the first time in their life. Among these children there will be many with very clear symptoms and a high level of disease activity, such that any test will find them, and many with no allergy and no symptoms, which are only tested because the parents are nervous. So the pediatrician will experience mainly children easy to diagnose and hence a higher accuracy than in the specialized allergy center setting. In other settings, the accuracy may be somewhere in between, for example for a general practitioner (GP) applying test in adults.

There have been also studies demonstrating empirically the dependency of the accuracy on the clinical population. Flicker et al. (1997) investigated the diagnostic accuracy of several dementia screening instruments for two different clinical populations. The first population was recruited from a specialised memory clinic, the second was a random selection of patients handled by an age care assessment team of a hospital. For the well known Mini-Mental Status Examination test (MMSE, Folstein et al. (1975)) they observed a sensitivity and specificity of 87.5% and 83.3%, respectively, in the second population, but only a sensitivity and specificity of 68.8% and 75.9%, respectively, in the first population. This difference probably reflects that in the second population many subjects had absolutely no signs of dementia (actually the prevalence was 32%), whereas in the first population most subjects were close to the border between dementia and non-dementia, as dementia is a continuous phenomenon (actually the prevalence was 72%).

With respect to the choice of the patient population for an accuracy study, we would like to distinguish three different steps in the sequel: The choice of the *target situation*, the choice of the *target population*, and finally the actual *study population* we can recruit for an accuracy study. If we are considering the accuracy of a diagnostic test, we typically have in mind a specific clinical situation where we intend to use the test for a specific purpose, i.e., to obtain a specific piece of information. This clinical situation is typically characterized by the fact that the patient has reached a certain step in the diagnostic or therapeutic process. We call this situation the *target situation*, and Table 2.1 shows broad categories of target situations. The target situation is typically characterized by clinical characteristics of the patient, which in particular describe how far the patient is in the current management process. For example, when we are interested in using an imagining technique to identify affected lymph nodes in patients with prostate cancer, we may define the target population by characteristics like

1. newly diagnosed prostate cancer
2. no bone metastases
3. scheduled for intended curative therapy
4. Gleason score > 6 and/or PSA concentration of 10 ng/mL and/or T3 cancer

(Poulsen et al., 2012). At first sight it might be desirable to investigate the accuracy of a test within all patients reaching the target situation. However, there may be very different subpopulations within this population, which are clinically relevant and for whom the accuracy may vary substantially. The example of the hazelnut allergy test mentioned above may illustrate this: the common target situation may be described as primary diagnosis in patients with a

Category	Characteristic of situation	Piece of information requested	Example
Population based screening	No restrictions based on symptoms	Disease yes/no	Mammography for breast cancer screening
High risk group screening	Persons have a high risk profile, but no symptoms	Disease yes/no	Enhanced screening in women with BRCA1 or BRCA2 mutations
Primary diagnosis	Patients with (new) symptoms, but no established diagnosis	Disease yes/no	Suspicion for appendicitis because of abdominal pain
Differential diagnosing	Patients with recently established disease	Subtype of disease	Subtypes of Parkinson disease
Staging	Patients with recently established disease	Disease Stage	Stages T1-T4 in lung cancer
Treatment planning	Patients with sufficiently characterized disease state to start treatment	Exact location to apply treatment	Radiation therapy
Response evaluation	Patients in whom treatment is started	Does the treatment work?	Decrease in tumor size/ activity after chemo therapy
Follow up	Patients treated successfully	Does the disease come back?	Follow up of cancer patients after therapy

Table 2.1: Broad categories for the target situation of a diagnostic test

suspicion for a hazelnut allergy, but we have already argued that the accuracy probably varies from setting to setting.

For a specific accuracy study, it can be wise to focus on one relevant target population which approaches the target situation in their clinical course. To find out what 'relevant' may mean here, we have to remember the goal of any accuracy study: We want to inform clinicians and patients about the accuracy they can expect if they apply the test/ the test is applied to them. This requires that the target population is homogeneous enough to justify the assumption that the accuracy we can observe in the target population is also applicable for each patient in the population. Mixing for example patients who have suffered from symptoms for many years with patients who developed symptoms recently is typically a poor idea, as the accuracy may substantially differ and there is a high risk of misinforming clinicians or patients. So duration or severity of symptoms are often useful characteristics to define the target population of a study. Sometimes a certain setting like general practice or emergency room can serve as the basic characteristic of the target population, if such a setting explains a lot of the heterogeneity in the difficulty to diagnose, and if the patients are rather homogeneous within the setting.

In the ideal case, the *study population* we can actually recruit for an accuracy study should be a random or at least 'representative' sample of the target population. However, the only approach to patients we typically have is via one or several centers which have contact to patients reaching or having reached the target situation. Consequently, the actual recruitment in each participating center actually determines the study population, and we have to consider how this recruitment may have an influence on the relation between the target population and the study population.

Nevertheless, it is sometimes rather simple to achieve the aim of a good agreement between study population and target population. For example if we are interested in determining the accuracy of sonography to diagnose appendicitis in general practice, it might be sufficient to recruit a sample of GPs who are willing to participate. It seems to be reasonable to assume that patients appearing with symptoms of an appendicitis do not differ substantially with respect to the difficulty to diagnose among different GPs. The situation is, however, different if we consider the application of an allergy test in general practice. Here we have to fear that the GPs' interest in allergies can widely differ, and that this interest is known to the patients, so that patients may select among different GPs in dependence on their disease history: Patients with a long history with inconclusive results may prefer to go to those GPs who are known for their interest and expertise in allergology. So depending on the choice of the GPs to be included in an accuracy study, the accuracy may actually differ. We can try to solve this problem either by narrowing the target population to those with specific characteristics of the patient history to reduce the population differences in difficulty to diagnose between the GPs, or we may decide

to consider two different target populations, defined by visiting a GP with specific interest in allergology or visiting a GP with normal interest.

Similar considerations may be necessary when using hospital departments for recruitment of patients. If patients reach a target situation typically in hospital, we have to check whether we can rely on that all patients reaching the situation in the hospital are really sent to this department, or whether they may be also sent to other departments. If the target situation can be reached outside of the hospital, we have additionally to check whether all patients are really sent to the hospital. Including only hospital departments which are the only care provider for patients in the local catchment area of a hospital are often the best source of recruitment, as then we can be pretty sure that the study population is close to the target population.

However, there is one additional key issue in avoiding undesirable differences between target population and study population: Within each center, all patients reaching the target situation and fulfilling the eligibility criteria for the target population must be included in the study. The most simple and convincing way to achieve this is to require that in a specific period all consecutive patients of the target population enter the accuracy study, and that both the index test and the reference test are applied. It would be dangerous to leave this decision to the treating clinician of each patient. Then it may happen that only patients with inconclusive results from the index test are sent to the reference test, such that at the end only patients difficult to diagnose are included in the study. It would be also dangerous to allow other subjects like the head of the department or the patient themselves to influence such decisions, as this may again prefer patients with high or low difficulty to diagnose to be included.

Remark: We included in Table 2.1 population based screening as one category. However, population based screening is a very special case compared to the other categories, as the prevalence of the target condition is typically very small. In the following, we will focus on the other categories.

2.4 Measures of Accuracy

The most common measures for the accuracy of a diagnostic test are sensitivity and specificity. Sensitivity is defined as the probability to obtain a positive test result for a patient for whom the target condition is present, and specificity is the probability to obtain a negative test result for a patient for whom the target condition is absent. Empirical estimates for these probabilities can be obtained by the corresponding relative frequencies.

Table 2.2 illustrates how sensitivity and specificity can be computed if the result of the diagnostic test in each patient has been classified into one of the four categories true positive

(TP), false positive (FP), true negative (TN), and false negative (FN). Here a positive result of the index test is called true positive, if the target condition is actually present in the patient – as indicated by the reference test – and false positive, if the target condition is actually not present. A negative result of the index test is called true negative, if the target condition is actually not present in the patient, and false negative, if the target condition is actually present.

		reference test		
		+	–	
index test	+	TP	FP	I+
	–	FN	TN	I–
		R+	R–	N

Table 2.2: Summarizing the results of an accuracy study in a four fold table and computation of estimates of sensitivity, specificity and predictive values. TP = number of true positive results, FP = number of false positive results, FN = number of false negative results, TN = number of true negative results, R+ = number of positive results of the reference test, R- = number of negative results of the reference test, I+ = number of positive results of the index test, I- = number of negative results of the index test, N = overall sample size.

$$\text{sens} = \frac{\text{TP}}{\text{R+}} \qquad \text{PPV} = \frac{\text{TP}}{\text{I+}}$$

$$\text{spec} = \frac{\text{TN}}{\text{R–}} \qquad \text{NPV} = \frac{\text{TN}}{\text{I–}}$$

Two further measures are the positive and negative predictive values. They express to which degree we can trust a positive or a negative test result of the index test, respectively. The positive predictive value (PPV) is defined as the probability that the target condition is present for a patient with a positive test result of the index test. The negative predictive value (NPV) is defined as the probability that the target condition is absent for a patient with a negative result of the index test. Estimates of these probabilities can be obtained by corresponding relative frequencies as illustrated in Table 2.2, too. In Chapter 8 we will extend this further.

2.5 Comparative Accuracy Studies

So far we have focused on the situation that we are interested in the accuracy of a single test. We have considered accuracy studies answering the simple question 'How good is the diagnostic test?' However, today we have for most target situations already some diagnostic tests. So

the typical research question today sounds 'Is the new diagnostic test better than the existing test(s)?' Consequently, we need to compare the accuracy of two or even more tests.

At first sight it might be obvious to approach such a question by comparing the accuracy of the different tests across different studies all investigating one test. And if we are interested in comparing a new test with existing tests we perform one additional study for the new test. However, such comparisons across studies are often not very convincing due to the difficulties to ensure that study populations match the target population (as we discussed in 2.3), or due to differences in the target population themselves. In particular, if study or target populations are mainly determined by care settings, comparability across countries is often questionable, as cultural differences or administrative structures may have a large impact on which patients approach which setting.

Consequently, to allow a meaningful comparison among diagnostic tests it is most convincing to conduct comparative studies where all tests of interest are applied in the same patient population, either by randomly selecting the test to be applied or by simultaneously applying all tests in each patient. (The choice between these two options is discussed further in Chapter 5.) Often, only two tests are compared: The new test and the current standard test of the clinical routine.

Most of what we have said so far about accuracy studies applies also to comparative accuracy studies. We have two index tests instead of one, and the final aim is to compare sensitivity, specificity and predictive values between the two tests. If both sensitivity and specificity improve by using the new test (which is equivalent to that both the positive and the negative predictive value improve), then we have a clear indication that the new test is better. In addition, statistical methods allow to assess whether such an improvement is beyond chance level. However, it often happens that we can observe an increase in sensitivity and a decrease in specificity, or vice versa. Then additional considerations are necessary to balance the two different types or errors (FP and FN decisions) against each other. We will come back to this point in Chapter 8.

Comparative studies can not only be performed if the new test should replace an existing test. They can also be performed if a new test should be applied in addition to the existing test. We have to compare the final test results with and without the new test.

One typical situation of this type is the so called 'add on' testing: We have an established diagnostic test, which we use for a treatment decision, i.e., only the test positive patients are treated. However, we are not satisfied with the positive predictive value of the existing test, i.e., too many patients who do not need treatment are actually treated. Consequently, we wish to improve the positive predictive value by applying one additional test only in the test positive patients, and to treat only patients who are also positive in this 'add on' test. So we have to compare the existing test with a new test, which is defined by combining the results of the

existing test with the additional test: We have a positive test result if and only if both the existing test and the additional test are positive (see Table 2.3). This comparison can again be performed by inspecting sensitivity, specificity and the predictive values of the existing test and the new one. We have only to be aware of that due to the specific construction of the new test the sensitivity can never increase (as fewer patients will experience a positive test result) and the specificity can never decrease (as more patients will get a negative test result). So it is here more natural to focus on the predictive values, which was also actually the motivation to introduce an 'add on' test. Another typical situation is often referred to as 'triage': Here

		Add on test	
		+	-
Existing test	+	+	-
	-	-	-

		Existing test	
		+	-
Triage test	+	+	-
	-	-	-

Table 2.3: Derivation of the results of the new test from the results of the add on test following the existing test (left) or from the results of the existing test following the triage test (right).

we have an established diagnostic test, but we are concerned about applying it in too many patients, as the test is invasive or expensive. We are interested in a simple and cheap test which we can apply prior to the established test in all patients, allowing us to continue with the established test in the test positive patients only. Again, we have then to compare the existing test with a new one based on combining both tests as illustrated in Table 2.3. As we have actually the same structure as in the 'add on' test, similar considerations can be applied.

Remark: Empirical evidence for the need of comparative studies instead of the comparison of single arm studies has been given in the paper by Takwoingi et al. (2013).

Remark: In this chapter we have considered dichotomous markers. Most we have said translates to quantitative (ordinal or continuous) markers. The difference is that when using a quantitative markers we either use the ROC curve as an analytic tool or we choose a cutoff to define a dichotomous decision rule. For these issues, see Chapter 8.

Summary of Chapter 2

To ensure that we can generalize the results of a diagnostic study, we need two basic prerequisites: First we have to be clear about the population in which we want to apply the test in the future (target population), and second we have to ensure that we can apply the test (and subsequent patient management) in clinical routine exactly like in the study. The key element in accuracy studies is the application of a gold standard test. Since real gold standards are rare, we typically use the term 'reference standard' to indicate that we use the best available method, which gives at least approximatively the truth. We can study the agreement between the index test, which is the diagnostic test we actually want to evaluate and the reference standard by classifying their results in a four fold table. From the latter numbers like sensitivity and specificity describing the accuracy of the index test can be computed.

Chapter 3

Benefit Studies

Objectives of Chapter 3

At the end of chapter 3 the reader should be able to ...

- understand the aim of a diagnostic benefit study
- recognize that benefit studies are typically performed as randomized trials
- name consequences of incorrect diagnostic decisions
- get a first idea about study design options of benefit studies

If a new diagnostic procedure improves the diagnostic accuracy, this is not in itself an advantage for patients. An advantage can only appear, if improved accuracy also leads to better treatment, or – generally speaking – to a better management of patients. The improved treatment then implies a better outcome, e.g., a prolonged survival. Only in this case we can conclude that patients benefit from improved accuracy. In a benefit study we try to address directly the question, whether a (new) diagnostic procedure implies really (in the long run) such an advantage (on average) for the patients compared to the current diagnostic standard. A typical example of a benefit study is a randomized trial, where half of the patients are randomized to obtain the standard diagnostic procedure, and the other half to obtain the new diagnostic procedure, and the further management and treatment of the patients is based on the results of the diagnostic procedure used. If we then observe that patients for whom the treatment decision is based on the new procedure have a better outcome, then we can be pretty sure that the new procedure is beneficial for the patients.

So the idea of benefit studies is quite simple, but – similar to accuracy studies – the actual design of a benefit study which is informative and convincing can be challenging. Again we have to ask how we can inform clinicians, patients and other stakeholders best about the benefit from applying a (new) diagnostic procedure we can expect if we would introduce it in clinical routine. And again this requires that the results of the study are generalizable to clinical routine, or that it can be at least judged to which degree they are generalizable. Not surprisingly, we arrive at similar key questions:

- What is the patient population we would like to make a statement about?
- How are the diagnostic procedures we compare exactly defined?
- How do the outcomes we compare relate to clinical practice and patient benefit?

Some of these questions can be addressed in a similar way as in accuracy studies. For example, our considerations about target situation, target population and choice of study population remain unchanged. However, some aspects change, as we are aiming at measuring benefit, not accuracy. Moreover, a fundamental difference relates to the fact that in benefit studies we actually evaluate a combination of a diagnostic test (or a more complicated diagnostic procedure) with some management decisions, i.e., a type of a complex intervention. We start this chapter by summarizing a lot of arguments why it is often necessary to regard diagnostic tests/procedures as part of complex interventions, which makes attempts to predict benefit from accuracy studies cumbersome. We next consider the choice of the outcome in benefit studies, which is of course a new aspect compared to accuracy studies. Next we explain why randomized trials are nearly always necessary in order to achieve an unbiased assessment of

the benefit. We then discuss some aspects of defining the interventions we actually compare in benefit studies, and finally introduce two designs for benefit studies which are useful in the early development of a new test. There is no specific section on the (statistical) analysis of benefit studies, as this follows essentially the established principles of analysing randomized intervention studies.

3.1 Viewing Diagnostic Tests as a Part of Complex Interventions

In some situations, the consequences of a correct or incorrect diagnostic decision may be so obvious, that there can be no doubt, that improved accuracy implies a benefit. However, this may be more the exception than the rule, due to different reasons which we discuss in the following.

1. Whenever we apply a diagnostic test, there are four different possibilities for the correctness of the test result, which substantially differ from each other with respect to the consequences for the patient.

    - The result can be true positive, i.e., the patient fulfills the target condition and the diagnostic test indicates this correctly. If the diagnostic test is the final one prior to the treatment decision, this implies that the patients will be treated and can benefit from the efficacy of the treatment. However, even 'effective' treatment options will often only imply a survival or cure rate of 50% or less. And this may depend on further characteristics of the patients like age, gender, disease stage or performance status. If the diagnostic test is not the final one, the positive test result has to be confirmed by other diagnostic procedures, and the benefit for the patients depends on the accuracy of these procedures.
    - The result can be false positive, i.e., the patient does not have the target condition, but the diagnostic test indicates this. If the target condition is 'diseased' and if the test is the final one prior to the treatment decision, the patient will get a therapy which is superfluous and futile. So the patient will not benefit from the therapy, but may suffer from its adverse effects. If the test is not the final one, the patient has to live at least for some time with an incorrect diagnosis, which may imply psychosocial consequences, including the possibility of suicide. If the diagnostic test aims to distinguish between two disease stages with different treatment options and

the target condition is the more advanced disease stage, the patient may have some benefit from the therapy indicated for the more advanced stage, but will suffer from not obtaining the most adequate therapy.

- The result can be true negative, i.e., the patient does not have the target condition and this is correctly indicated by the diagnostic test. Then the consequences are typically positive for the patient. The patient can avoid any therapy (and feel happy) or will receive the therapy appropriate for his or her disease stage. However, if the patient is suffering from (severe) health problems, a negative test results may actually imply additional burden, as the search for the cause of the problems continues.
- The result can be false negative, i.e., the patient does have the target condition, but the diagnostic test indicates incorrectly its absence. If the negative result implies no further testing then the patient typically does not get an adequate therapy. For example when the target condition is a non-local disease state in a cancer patient, a false negative result implies the application of a local radiation therapy or a local surgery, instead of a therapy targeting also metastases like a chemotherapy. In general, the actual impact depends often on when the presence of the target condition will be detected in the follow up of the patients and whether it will be then too late for an effective treatment or not. In addition, the false negative result may by itself imply a prolonged delay in detecting the target condition, as the patient may be convinced to lack this condition, and hence he or she may react less sensitive to symptoms or other signs.

So the overall benefit for the patients of a single diagnostic test is a mixture of the consequences for each of the four situations where each contributes with a weight according to the accuracy of the test. And the consequences for each situation may be unclear or unpredictable. The situation is even more complicated, if we compare two diagnostic tests, as then the overall benefit depends on the consequences of changing the diagnostic decision. We will come back to this in Chapter 7.

2. Treatment options may be less or more effective than we believe, because a new diagnostic test with higher sensitivity actually adds a new subgroup of patients. For example, with a new diagnostic test we may detect patients in an earlier disease stage than before, and a treatment may be more effective in this subgroup than in other patients. However, it may also happen that we detect a new subgroup which does not benefit from the typical treatment. For example, new screening tools for depression may detect patients with a very mild form who do not need any pharmacological treatment.

3. We cannot be sure whether treatment and management options are chosen in the way we may believe or wish. For example, clinicians may interpret a positive test result from a new diagnostic procedure differently than the 'same' result from the current standard. They may regard it as more trustworthy than a positive result from the standard procedure, and hence omit additional diagnostic procedures they requested when using the old standard. They may regard it as less trustworthy, which may result in the extreme case that they always request also the old standard procedure and then rely more on the results of the old one than the new one.

4. Diagnostic tests may provide additional information beyond that on the target condition, which may have an impact on patient management. For example, when using an imaging modality to decide whether a tumor is resectable or not, a positive test result implies that surgery is performed, but the image may change the surgery itself. Even if the modality improves the decision on resectability, we cannot be sure that the surgical procedure is still beneficial, as the image may misguide the surgeon, for example the tumor is not removed completely. Similar, if imaging procedures are used to judge a certain body region, they may provide additional information on neighboring regions. This information may not be ignored by the treating physician.

5. A diagnostic procedure may be investigated in accuracy studies at another level than relevant for a benefit at the patient level. Accuracy studies may for example suggest that a diagnostic procedure has a high accuracy in detecting local metastases, if it is analyzed at the level of single metastases. However, for an effective treatment – e.g., by surgery – it may be necessary to know all local metastases. Even if the accuracy is high at the lesion level (i.e., for each single local metastasis), the procedure may fail to detect very small metastases. It may be that nearly all patients with metastases have at least also one small metastasis, which we overlook, and hence there will be no benefit from the surgery.

6. A diagnostic procedure is applied several times, for example as part of the follow up monitoring of cancer patients after therapy. Then the benefit does not only depend on the accuracy, but also on the spacing of the visits, the compliance of the patients, and the intra-individual differences in the speed of tumor growth.

All these reasons and many similar ones suggest that when asking the question about the benefit of a diagnostic test, we have to regard this test as part of a complex intervention, which consists at least of one additional component, namely a management or treatment decision. However,

often it is a whole management and treatment process we start with a diagnostic procedure and we have always at least two completely different types of processes, namely one for the test positive and one for the test negative subjects. There can be many different issues in those processes – foreseeable and unforeseeable ones – which can have an impact on the final benefit for the patients.

3.2 Choice of Outcome

If we use the term 'benefit', we are implicitly taking a patient perspective. This means, we are talking about an advantage for patients, at least on average. So the (primary) outcome variable should be chosen such that an improvement in this outcome reflects a true advantage for the patient. Such outcomes are called 'patient-relevant outcomes', 'patient-centered outcomes', or 'patient related outcomes'. The triad 'mortality, morbidity and quality of life' are often regarded as the most prominent examples of such outcomes. The choice of the outcome has typically little to do with the diagnostic procedures we investigate, but with the clinical problem and treatment strategies. Thus one should choose the same patient-centered outcomes as in corresponding treatment studies. In the case of potentially lethal diseases overall survival is typically the best choice. Morbidity measures like blood pressure or disease progression are sometimes problematic, as their direct impact for the patient is questionable. A patient-relevant outcome is the occurrence of any serious adverse events during or after the treatment. Quality of life measures are appropriate if a disease mainly effects the quality of life, e.g., due to pain or fatigue.

There are two types of outcomes, which are definitely not appropriate when investigating the clinical benefit of diagnostic procedures. The first type of outcomes is given by those which reflect the accuracy, for example the number of correct diagnoses or the number of correct changes of the diagnosis. Here we do not get any more information than in accuracy studies, and we do not know the long term impact on the patients. The second type of outcomes is focusing on management decisions, e.g., the number of treatment decisions changed or the number of surgeries initiated. Such outcomes do not allow us to make conclusions on the benefit, as we do not know whether the decisions are beneficial for the patients or not. However studies with management decision as outcome can be useful as a step between accuracy studies and benefit studies, because if a (new) diagnostic procedure has no impact on management decisions, it can neither have a clinical benefit (see Chapter 4).

As an example how difficult it may be to judge whether an outcome is patient important or not, we may consider the outcome 'number of futile thoracotomies', as used for example

by van Tinteren et al. (2002) and Fischer et al. (2009) in evaluating the benefit of PET/CT in managing patients with non-small cell lung cancer. There can be little doubt that avoiding futile surgeries reflects an advantage for a patient. However, when counting the number of futile thoracotomies performed, we automatically interpret any patient without a thoracotomy as a futile thoracotomy avoided. But this means that we have to rely on that the management decision not to perform a thoracotomy was always correct. This can be only justified if we have some type of external gold standard for these patients, allowing us to validate this management decision.

3.3 The Need for Randomized Trials

Once we have agreed on that we need a specific study to assess the benefit of a (new) diagnostic procedure, and once we have agreed on a relevant and valid outcome to assess the potential benefit, it remains still the question of the adequate study design. New diagnostic procedures are often implemented in clinical routine in a rather informal process, as – in contrast to pharmacological treatments – there is no elaborated regulatory process. This may suggest using observational data to study benefit, for example by simply comparing patients who underwent the new procedure with those who underwent the old procedure in the same hospital. There are, however, typically some issues which prevent this simple idea from generating a valid and useful comparison:

- The two patient groups may not be identical with respect to their prognosis. The new procedure may be mainly applied in patients at risk for an advanced disease stage, or when the results of the standard procedure were inconclusive.
- The two procedures may be offered by two different departments, e.g., CT by the radiology department and PET by the nuclear medicine department. The patient populations of the departments or the management strategies may differ.
- In many patients both the standard procedure and the new procedure are applied, and it remains unclear, how the results have determined the management process of the patients.
- Relevant outcome data may not be collected as a part of clinical routine, its validity may be questionable, or the follow-up for the patients undergoing the new procedure is more intensive.

Instead of making a comparison within a hospital, one may prefer to compare hospitals who use the new procedure with hospitals using the old one in clinical routine. Then some issues are

vanishing, as the choice of the test is determined externally. However, the comparison is then confounded with differences between the hospitals. The hospitals applying the new procedure may be also those with better management strategies or attracting more high risk patients.

At the end, there is hence often a need for a prospective, randomized control trial, where patients are randomly assigned to undergo the old or the new diagnostic procedure. This way we can solve many of the issues mentioned above:

- The two patients groups are not systematically differing with respect to their prognosis and the degree of diagnostic difficulty
- The follow-up of the patients and measurement of the outcomes can be organized in an identical manner in the two patient groups.
- We can – at least to some degree – ensure that patients undergo only one of the two diagnostic procedures to be compared.

3.4 What Do we Evaluate in a Diagnostic Benefit Study?

The aim of a diagnostic benefit study is to inform clinicians, patients, guideline developers and policy makers about the benefit we can expect for the patients if we use one diagnostic procedure instead of another one in clinical routine. To approach this aim, we have to be clear about what we exactly evaluate in a diagnostic benefit study and to which degree and under which circumstances we can generalize the results to the real world outside of the study.

There are two key issues: First, we have to define and describe the patient population to be included as precise as possible, to ensure that the 'better' diagnostic procedure is in future really used in those patients, for whom it is better. With respect to the choice of the population, the same considerations as for accuracy studies apply: We have to be clear about the target situation and the target population and have to take into account how the recruitment may lead to a study population different from the target population (see 2.3). Second, we have to define and describe the two interventions as precisely as possible. Here we have to be aware of that we compare two (complex) interventions, consisting of the application of a diagnostic test (or a more complex diagnostic procedure) and a subsequent decision and management process. All these components have to be clearly defined and described. With respect to the diagnostic test we can refer to our considerations in 2.2. As benefit studies are often conducted as multi-center trials, the question of the adequate level of standardization may come now further into the focus. A full standardization may not always be desirable. In the real world diagnostic procedures will be always used in a somewhat non-standardized manner. For example different

hospitals will use different machines for the same imaging procedure, and different labs will use different assays. In multi-center studies a high degree of standardization may imply an artificial homogeneity, as this homogeneity will be never met in reality. Often common training based on a standardized and published manual is a good option to reach a sufficient and generalizable degree of standardization.

The additional step in benefit studies is to define and describe the processes taking place after a diagnostic test has been applied. Sometimes, this task is very simple and easy. If the diagnostic test should distinguish between two disease states, if it always results in a clear decision, and if there exist two widely accepted and evidence based unique treatment recommendations, one for each disease state, then it is no problem to define and describe a highly standardized decision process and the treatment: In dependence on the result of the test, the recommended treatment is given. However, standardization of the treatment is often a more subtle issue, as even if one agrees on the disease state, the treatment may not be exactly identical. For example the exact performance of a surgical treatment option may differ from hospital to hospital due to different traditions, or from clinician to clinician as it is based on the clinician's individual experience. Another complicating factor may be that the treatment may not only depend on the clinical condition but also on patient characteristics like age, performance status or the results of other diagnostic tests.

Any treatment and management decision process can be viewed as an algorithm, and the task is to define and describe this algorithm with sufficient precision. Moreover, it might be necessary not only to define the actual treatment, but also the long term management, for example a schedule for follow up visits with specified clinical investigations. For example, if a diagnostic procedure is used for screening in high risk population, the benefit for the patients is often highly depending on the screening interval. The task becomes even more complicated, if the diagnostic procedure provides a more complex information like a simple yes/no decision, for example a whole body image which is used for several decisions or as a guidance for a surgery.

In general, we have the same challenge as with the diagnostic procedure: On the one hand, we want to standardize the management process including treatment decisions and treatments as much as possible, such that we can ensure that patients get the intended treatments making full use of the diagnostic information available, and such that we can transfer the results of the study to routine care. On the other hand, we have to take into account that also in routine care there will be later some heterogeneity, which we cannot avoid. So the task is to achieve a reasonable degree of standardization without introducing artificial homogeneity. There is, however, often a fundamental difference between the possibility to standardize a diagnostic test and to standardize the following management process. Diagnostic benefit studies are often initiated and planned by those health professionals, who are in charge for the diagnostic task.

The common interest is then often sufficient to obtain a reasonable standardization also in a multi-center setting. However, treatment decisions and treatments are often performed by other health professionals from other departments or even from other sections of the health care system. The interest in standardization is often less pronounced in these health professionals, as the study is not of direct interest for them. So this may put a practical limit on the possibility to standardize management.

Finally, we should mention a further consequence of actually evaluating a combination of a diagnostic procedure with a treatment decision and management process: if a randomized benefit study fails to demonstrate a benefit, this does not imply that the diagnostic procedure is worthless. The failure may be due to the fact that the treatments used were not effective enough, or that the treatment decision process was not optimal. So whenever we start a randomized benefit study, we should be pretty sure, that all management decisions are optimal and that the different treatments we offer the patients make really a difference. (In Section 9.2 later we will discuss how big this difference should be.) And it is often questionable whether we can be sure about this, as a new diagnostic procedure often does not only make more accurate decisions, but actually identifies a new subgroup of patients overlooked by the old one, and this subgroup may be special. For example, in cancer patients, new imaging procedures may help to detect patients with small metastases. However patients with (only) small metastases may benefit from a curative treatment, which we today only offer to patients with no metastases. So offering these patients now only a palliative treatment instead of a curative treatment – as indicated by the new diagnostic procedure – and forgetting to distinguish between small and large metastases may imply really harm to patients!

Whenever there is some doubt about whether patients really benefit from the treatment options we have in mind to offer in dependence on the results of a diagnostic procedure it may be too early to start directly a benefit study and the alternatives considered in the next section may be more adequate.

3.5 If we are in Doubt about Optimal Treatment

There may be some doubt about how to use this information for the further management of the patients, in particular if a diagnostic procedure provides a new type of information which was not available previously, This is for example the case when introducing a new biomarker, and we have the idea that the biomarker positive patients are benefitting from a specific therapy or a specific add-on to the current therapy, which takes advantages of the specific condition indicated by the biomarker. This idea may be grounded in knowledge about biological pathways,

but a proof from an RCT is missing. Another example may appear if a new diagnostic procedure allows us to identify patients at an intermediate risk level between two established disease states, and it is unknown which of the two therapies offered today in dependence on the disease state is appropriate for this group. Patients with small metastases mentioned in the last section build an example of this type: can they benefit from a curative treatment like patients without metastases, or is a palliative treatment more appropriate like for patients with large metastases.

If we are in doubt about the choice between two treatment options for the subgroup of patients identified by the new diagnostic procedure or by a specific discrepancy between a new and an established diagnostic procedure, it is a straight forward idea to perform a therapeutic RCT, randomizing the patients to the two treatment options, but including only the subgroup of patients we are in doubt about. So this may be for example biomarker positive patients, or the patients with small metastases identified by the new procedure and overlooked by the old. We refer to studies following this idea as preselection designs, and we will discuss them further in Section 6.3.

It may also happen that we are in doubt about the value of two treatment options A and B for all patients, but that we may have the idea that a diagnostic procedure can help to identify those patients benefitting more from A than from B and vice versa. Then it is a straightforward idea to apply the diagnostic procedure to all patients and to randomize all patients to A or B, and then to study how the treatment difference between A and B varies in dependence on the results of the diagnostic procedure. In treatment research, such studies would be regarded as an attempt to establish the diagnostic procedure as a 'predictive factor'. In diagnostic research, they are known as 'interaction studies'. In the most simple case, the diagnostic procedure may result in two possible states, 'positive' and 'negative', and the interaction study may allow to demonstrate that A is better than B in positive, but B is better than A in negative patients.

Further reading: General discussions about the planning of benefit studies can be also found in the papers by Gazelle et al. (2011) and by Ferrante di Ruffano et al. (2012b).

Summary of Chapter 3

Benefit studies are typically performed as randomized trials with patient relevant outcomes like survival. The simple idea is to randomize the patients to two different diagnostic tests we would like to compare (e.g., two different imaging techniques like ultrasound and computer tomography), and then to compare the long term outcome between the groups.

Chapter 4

Phases of Diagnostic Research

Objectives of Chapter 4

At the end of chapter 4 the reader should be able to ...

- recognize that accuracy and benefit studies are only pieces in the whole process of developing a diagnostic test
- recognize that different kinds of studies are required at different stages of research

Accuracy and benefit studies are the most prominent studies in diagnostic research. However, there are also other types of studies in diagnostic research. They are not considered in more detail in this course, but it is good to know them. In particular they reflect that accuracy and benefit studies are only some (important) pieces in the whole process of developing a diagnostic test. We describe in the following the ideal process (Lord et al., 2006).

At the start of any new diagnostic test there is an idea about a new approach to get a certain piece of information about the disease state of a patient. This idea may be very simple, for example a simple question to the patient. Then the test is given by the idea itself. But typically the idea requires developing an instrument we will actually use in the test. This instrument can be a lab test, an imaging procedure, a questionnaire, a symptom list or something else, or a modification of an existing instrument of these types. The developing of such an instrument may already involve some type of small studies, for example experimental studies to optimize a lab test, or the psychometric validation of a questionnaire.

Whenever there is a subjective component in the instrument or in the test, it is wise to check that the impact of this subjectivity on the test results is of limited degree. Otherwise, we cannot expect a high accuracy under clinical routine conditions, where different subjects will apply the test (cf. 2.2 and 3.4). Here reproducibility studies are the most appropriate approach: we check, whether different subjects come to the same test results in a series of patients (inter observer variability) and whether same the subject comes to the same test results in a series of patients (intra observer variability). The analysis of such studies can make use of statistical methods like limits of agreement and the intra class correlation coefficient (in the case of continuous outcomes) or Cohen's kappa and related statistics (in the case of binary or categorical outcomes). If instruments with many items are used, psychometric methods can be used. Reproducibility studies can not only be used to demonstrate a sufficient reproducibility, but also to optimize the conditions for obtaining sufficient reproducibility, e.g., by varying the amount or type of training between observers.

If we believe that the instrument is optimized, we can start to investigate its accuracy. If an established test already exists to obtain the piece of information of interest, it is wise to start from the beginning with comparative studies comparing the new test with the established standard test. If a test can be applied in different target populations, it will be natural to perform accuracy studies for any target population of interest. It is very useful when several research groups perform an accuracy study for the same target population to investigate the robustness of the results. Multiple accuracy studies for the same target population can be also used to optimize the accuracy, similar as we have mentioned it for reproducibility studies.

Once sufficient accuracy has been established, we can start to try to approach the question of benefit for the patients. If the test provides a new piece of information which we intend to use

for a treatment decision, interaction studies or preselection designs may be the next step. If the test provides an established piece of information, it can be of interest to study whether the new test really results in the expected change of management, if the result differs from the standard test. This question can be approached in a prospective study similar to an accuracy test: In the target population of interest the established standard test and the new test are applied, but first only the result of the standard test is communicated to the treating physician, and he or she has to make a management decision. Then also the result of the new test is communicated, and the treating physician is allowed to make a new decision. The primary outcome is then the number of changes in management, or the number of changes in management which are in accordance with current treatment guidelines. If the new test is replacing the standard test in some hospitals as a consequence of the results of the accuracy studies, it is also possible to compare the distribution of management decisions before and after introducing the new test, if we expect a movement into one direction. However, then some of the limitations discussed in 3.3 may apply.

If we are sure that we know the optimal way of using the information from a (new) diagnostic test and that we can standardize the management process accordingly, we may start to think about performing randomized benefit studies as described in 3.

However, randomized benefit studies are still not the end of the story. If benefit studies have led to a decision to introduce a new diagnostic test or a more complex diagnostic procedure in clinical routine, it might be still wise to make an additional check, whether this decision has resulted in the benefit all the studies have promised. So we may check that we really detect now more of the diseased patients and at earlier stages, that we offer more often a treatment which is adequate, or that the quality of life or the survival of the patients has really improved. We should also check whether the new test has really replaced the old test, or whether it might be used in addition.

This general idea about the process of developing a diagnostic test has led to several suggestions to define phases of diagnostic research. A popular suggestion is due to Fryback and Thornbury (1991), who distinguish 6 levels: technical efficacy – diagnostic accuracy efficacy – diagnostic thinking efficacy – therapeutic efficacy – patient outcome efficacy – societal efficacy. A similar suggestion has been made by Kobberling et al. (1990).

For the area of diagnostic tests based on biomarkers, Pepe et al. (2001) suggested a somewhat different scheme. This reflects the fact that tests based on biomarkers can be often applied without physical presence of the patient, by using stored tissue or blood samples. Hence retrospective analyses of prospectively collected bio samples play a more central role in this area of research.

Further reading: The paper by Sackett and Haynes (2002) entitled "The architecture of diagnostic research" gives a nice summary about the expectation we should have in different steps of developing a diagnostic test.

Summary of Chapter 4

Different kinds of studies are required at different stages of research. Whereas experimental studies can be an instrument to optimize a diagnostic test, subjective components of a test can be studied through reproducibility studies. Comparative accuracy studies are conducted to investigate the accuracy of a diagnostic test. After the establishment of sufficient accuracy, RCTs are a useful tool for analyzing the benefit of a new test.

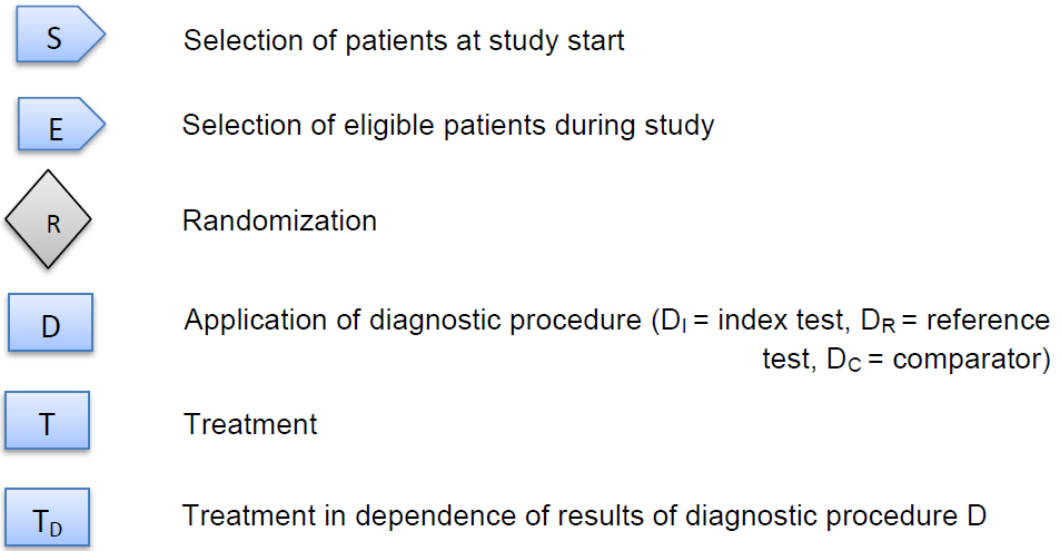
Part B

Design Options in Diagnostic Research

Introduction Part B

In part A we have considered some basic issues in planning diagnostic accuracy or benefit studies. In part B we give an overview about actual design options for conducting diagnostic accuracy or benefit studies.

In describing the design options we use a common scheme. We start with a description of the key property of the design, with a flow diagram illustrating the patient flow in a study according to this design. Next we describe the research question we can actually address with this design. This research question typically reflects a certain step in the process of developing a new diagnostic test or procedure. Subsequently, we give a more detailed description of the design, followed by a short description of the analytical strategy typically applied to this design. We emphasize that the names for the designs are not always used consistently in literature, and thus refer to different names used for this design. Finally, we present an example from the medical literature where the design has been used. In the flow diagrams we use the following symbols:



Chapter 5

Design Options for Accuracy Studies

Objectives of Chapter 5

At the end of chapter 5 the reader should be able to ...

- recognize that there are various design options for conducting accuracy studies
- differentiate between single arm and comparative accuracy studies
- differentiate between prospective and case-control designs

In this chapter, we introduce varoius design options for diagnostic accuracy studies. There are prospective single arm accuracy studies, case-control accuracy studies, paired comparative accuracy studies, and randomized comparative accuracy studies.

5.1 Prospective Single Arm Accuracy Study

Key property

Patients are selected to the study from a relevant target population based on the suspicion of the target condition. All patients eligible are tested with the new test and the reference standard test (figure 5.2).

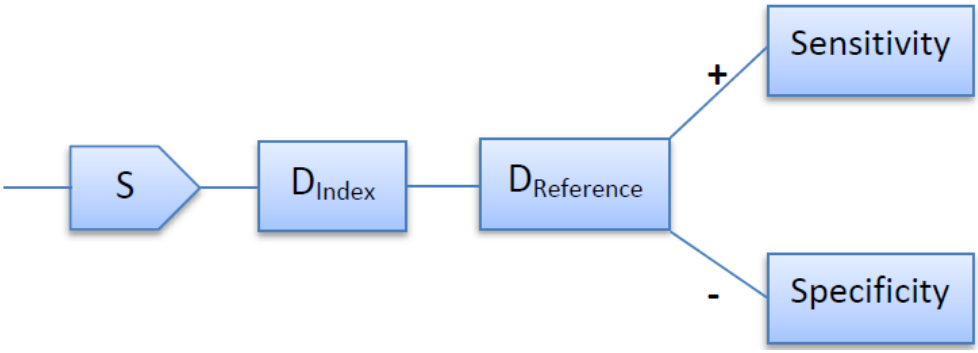


Figure 5.1: Prospective single arm accuracy study

Research question

We have developed a new test and now we want to know whether the new test is 'good enough'. Thus, the main question is: 'How good is the test in correctly classifying patients with respect to having or not having the target condition'. For this reason we have to demonstrate sufficient accuracy of the new test. This design allows to study the accuracy of a single test and there is no comparison with other existing tests. If a comparison with another existing test is intended a comparative accuracy study has to be conducted (see Section 5.3).

Description

All patients who approach the health care system in the target situation and belonging to the target population are included in the study. Typically the situation is given by a suspicion about the target condition, which is based on symptoms or the results of other diagnostic tests. The index test, i.e., the new test of interest, is performed in all patients, and the presence of the target condition is determined by performing the reference test in all patients, too. In principle,

the accuracy remains the same regardless whether the index test or the reference standard test is performed first. In practice, typically the new test is performed first and the reference test afterwards. It is essential that tests are performed 'blind', that means, that the reference standard has to be applied and interpreted in total ignorance of the test result of the index test and vice versa.

Analytic strategy

The accuracy of the index test is studied by comparing the results of the index test with the results of the reference standard. The results can be summarized in a 2×2 table and accuracy can be expressed as sensitivity, specificity, and predictive values of the test (cf. Section 2.4).

Variants

Reversed-flow design (Rutjes et al., 2005): For this variant the reference standard test is applied first to the patients. Then the index test is performed for all patients, i.e., for both those with and without the target condition. The authors themselves classify this design as a case-control design, because the disease state of the patient is known before the index test is performed. Considering the fact that the order of index and reference standard test does not influence the accuracy, we think it is justified to regard this design as a variant of a prospective single arm accuracy study if the assessor is blinded with respect to the patient's reference status.

The special case of using a reference standard based on follow-up information has been named *Delayed type cross-sectional study* (Knottnerus and Muris, 2003).

Other names and references

In the literature there are many other names for this type of study design, for example:

Classical design (Rutjes et al., 2005)

Single diagnostic test evaluation (Moons et al., 1999)

Survey of total study population (Knottnerus and Muris, 2003)

Cohort type accuracy studies or single-gate studies (Bossuyt and Leeflang, 2008)

Cohort studies (Pepe, 2003)

The name 'cohort study' refers to the fact that the design is similar to a prospective cohort study in epidemiology, where patients are also enrolled consecutively at different time points.

Example: Detection of bladder tumors

Golijanin et al. (1995) investigated the usefulness of immunostaining of the Lewis X antigen in cells from voided urine for the detection of bladder tumors. Cystoscopy followed by biopsy was considered as the reference standard. This approach is invasive, associated with complications, and expensive. Hence there is a need for noninvasive and cheap alternatives.

In their abstract Golijanin et al. (1995) described the basic set up of their study in the following way: 'Three consecutive voided urine specimens were obtained from 101 patients, 78 of whom were under surveillance because of a history of bladder tumors, and 23 were being evaluated because of hematuria or irritative urinary symptoms. Indirect immunoperoxidase staining of two urine samples was done on cytocentrifuge slides, using the P12 monoclonal antibody against the Lewis X antigen. The diagnosis of the presence of a urothelial tumor was made if more than 5% of the cells showed a typical red-brown staining. Cytopathologic examination of the third urine specimen was done according to Papanicolaou. Each patient underwent cystoscopy, and biopsies were obtained whenever there was endoscopic evidence of bladder tumors or carcinoma in situ.'

Analysing the two samples of each patient separately (i.e., considering 202 tests performed in 101 patients), a sensitivity of 81% could be reached. Analysing the two samples together and requiring at least one sample to be positive to indicate that a patient has a tumor increases the sensitivity to 97% with a specificity of 85.5%.

5.2 Case-Control Accuracy Study

Key property

Two different sets of patients are recruited: patients with verified presence of the target condition (cases) and patients without the target condition (controls). The index test is applied in both groups. The sensitivity is calculated in patients with the target condition and specificity in patients without the target condition. No reference test is applied (figure 5.2).

Research question

The research question is the same as in a prospective single arm accuracy study.

Description

Usually, cases and controls are sampled from two different sources. Individuals with verified target condition are typically sampled from a clinical population, for example a hospital. Controls are sampled from another population. The name 'two-gate design' describes that individuals

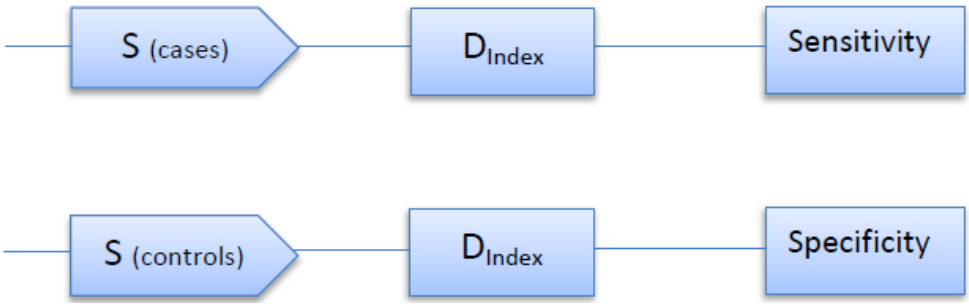


Figure 5.2: case-control accuracy study

enter the study through two separate gates entrances, using different inclusion criteria for cases and controls.

The choice of controls is a crucial issue in this design. Depending on the phase of test development, the objective of a diagnostic accuracy study can vary and with it the choice of cases and controls. In an early evaluation of a new test it can be convenient to use healthy controls to learn about the potential of the test. In a more advanced phase, healthy controls from the general population are not useful, as they do not represent the clinical population of interest. Then, cases and controls have to be chosen closely from a clinical setting in which the test is intended to be applied (Sox et al., 1989). This design is in particular popular in biomarker research, selecting cases and controls from bio banks.

Analytical strategy

The sensitivity is calculated in patients with the target condition and specificity in patients without the target condition.

Variants *Two-gate design using alternative diagnosis controls* (Rutjes et al., 2005): A different form of two-gate sampling includes only control participants diagnosed with a specific alternative condition known to produce symptoms and signs similar to those of participants with the target condition.

Other names and references

Two-gate design (Rutjes et al., 2005)

Case-referent approach (Knottnerus and Muris, 2003)

Example: Imaging tests in young women with breast symptoms

Houssami et al. (2003) investigated the value of mammography in young women with breast symptoms. 480 women were considered who went to a symptomatic breast clinic, older than 25 and younger than 55. 240 women with breast cancer constituted the 'cases' and 240 patients without breast cancer constituted the 'controls'. In both groups, cases and controls, mammography was performed. Sensitivity and specificity were investigated for the whole population and for age groups ($\leq$35, 36-40, 41-45, 46-50, 51-55).

To calculate the sensitivity of mammography the 240 cases were considered. Thereof 182 women had a positive result after mammography which leads to a sensitivity of 75.8%. In the controls, a specificity of 87.6% could be observed. Consideration of the different age groups shows that there seems to be an improvement in sensitivity for older patients whereas there seems to be no dependence between age and specificity.

5.3 Paired Comparative Accuracy Study

Key property

Patients are selected to the study from a relevant target population based on the suspicion of the target condition. All patients eligible are tested with the existing test (comparator), the new test and the reference standard (figure 5.3).

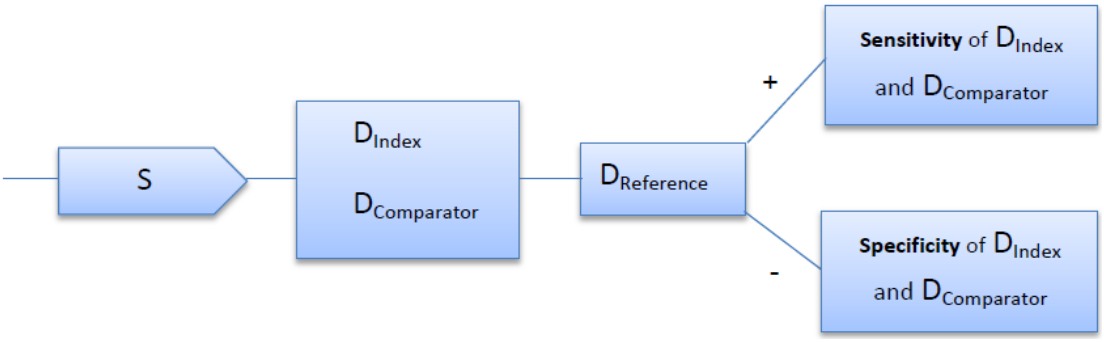


Figure 5.3: Paired comparative accuracy study

Research question

We have developed a new test and we want to compare its accuracy with that of another existing test (comparator). The comparator is typically the existing standard test, i.e., the test which is used in current practice. So, we are interested in the diagnostic accuracy of the new test relative to the diagnostic accuracy of the comparator.

Description

A consecutive series of patients suspected for having the target condition are tested with the existing test (comparator) and with the new test. All patients undergo also the reference standard test. The accuracy of the new test and of the comparator will be assessed by comparing the results with the results of the reference standard. All three tests have to be performed blinded for the results of the other tests.

Analytical strategy

The accuracy of the new test and of the existing test are compared within individuals, using statistical methods that acount for the paired structure of the design. If the new test shows higher sensitivity and specificity, it is the better one. If one parameter is improved, but the other becomes worse, sensitivity and specificity have to be balanced against each other. Confidence intervals for the change in sensitivity and the change in specificity can support the final decision.

Variants

Paired comparative accuracy studies can also be conducted as case-control studies, as shown in Figure 5.4.

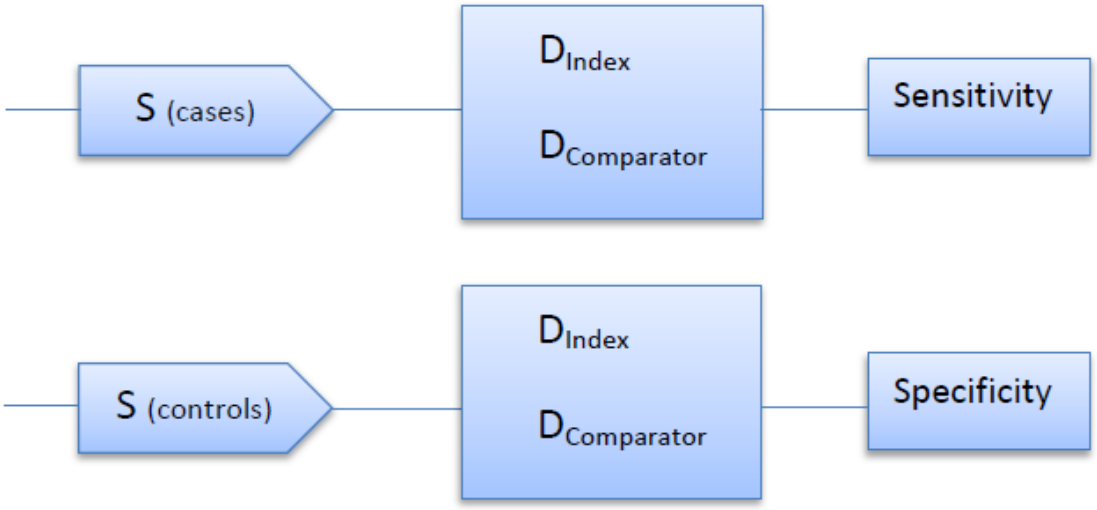


Figure 5.4: Paired comparative accuracy study in case–control design

Other names and references

Paired comparative accuracy study (Bossuyt and Leeflang, 2008)

Paired or cross-over comparative accuracy study (Takwoingi et al., 2013)

Example 1: Diagnosis of Hirschsprung's disease

De Lorijn et al. (2005) conducted a prospective comparative accuracy study concerning the diagnosis of Hirschsprung's disease (HD), which is a rare reason for constipation in infants and children. In the study the accuracy of three tests were compared: contrast enema (CE), anorectal manometry (ARM) and rectal suction biopsy (RSB). In each participant with suspicion of HD all three tests were performed. If there were two or more positive test results after the application of the index tests or if the bowel complaints continued, the reference standard was a full thickness biopsy to verify the absence of ganglion cells which indicates HD. That is, the reference standard depended on the index tests. Otherwise the reference standard was clinical follow-up (minimum 6 months).

122 participants were included into the study. In 111 of them all three tests were performed in arbitrary order within three weeks. In 28 children HD was diagnosed. Each of these children had two or more positive results after application of the index tests. In each test there have been inconclusive cases (8 CE, 15 ARM, 2 RSB) which were excluded in the calculation of sensitivity and specificity. Considering the CE test 19 of the 28 children with HD had a positive test result. For 6 diseased children the CE test was negative and for the remaining three diseased children the test was inconclusive. This leads to a sensitivity of 76% (CI: 57%-89%). The specificity amounts 97% (CI: 91%-99%). The ARM test detected HD in 19 of the 28 children with HD and without the excluded 15 children with inconclusive results there is a sensitivity of 83% (CI: 63%-93%) and a specificity of 93% (CI: 85%-97%). The RSB test detected the most cases of HD. 25 of the 28 children with HD got a positive test result. No child without HD got a positive test result. The sensitivity is 93% (CI: 77%-98%) and the specificity is 100% (CI: 96%-100%).

With these results RSB seems to be the most accurate test. The values of sensitivity and specificity were highest for RSB. However, they were not significantly different from the corresponding values of the other tests. Moreover, RSB produced the lowest number of inconclusive results.

Example 2: Imaging tests in young women with breast symptoms

Recall the study of Houssami et al. (2003) from the example in section 5.2. Actually, this was a comparative, paired case-control study, as sonography was applied in addition to mammography in all women.

Comparing the two tests there was a difference of 5.9 (95% CI -1.5% - 13.2%) for sensitivity (81.7% sonography, 75.8% mammography). Considering the results concerning specificity there was a difference of 0.4 (95% CI -5.0% - 5.8%) (88.0% sonography, 87.6% mammography).

Also the two age groups 'younger than 45' and 'older than 45' were considered. In the

younger group the sensitivity of sonography was 84.9% compared to 71.7% for mammography. This is a difference of 13.2% (95% CI 2.1% - 24.3%). In contrast, for the older women the same sensitivity could be observed for sonography and mammography (79.1%).

5.4 Randomized Comparative Accuracy Study

Key property

Patients are selected to the study from a relevant target population based on the suspicion of the target condition. All patients eligible are randomized either to the existing test (comparator) or to the new test, but the reference standard is applied to all of them (figure 5.5).

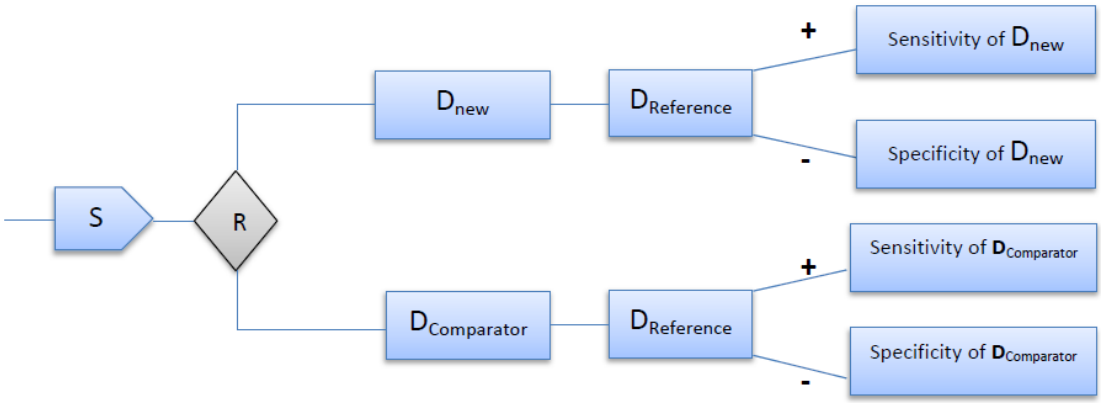


Figure 5.5: Randomized comparative accuracy study

Research question

The same as in a paired comparative accuracy study, but the application of both tests in each patient cannot be justified. Typical reasons are:

- The tests are too invasive for the old and new test to be done in the same patient.
- The tests interfere with each other.
- The study has additional objectives, such as assessing adverse events.

Description

Patients suspected for the target condition are randomly allocated to be exposed to either the comparator test or the new test. Additionally all patients undergo the reference test.

Analytical strategy

The same as in paired comparative studies. However, statistical methods to compute p-values and confidence intervals differ.

Other names and references

Randomized accuracy study (Takwoingi et al., 2013)

Example: Abdominal and iliac arterial stenoses

Schaefer et al. (2006) performed a comparative randomized study considering the detection of abdominal and iliac arterial stenosis. The aim was to compare three-dimensional magnetic resonance (3DMR) angiography with either gadodiamide or gadopentetate as contrast agent. As reference standard intraarterial digital subtraction angiography (DSA) was used. Stenosis $\geq$50% was defined as the target condition. 247 participants were randomized either to the gadodiamide or the gadopentetate group. The reference standard DSA was performed in both groups, either two weeks before or after the respective 3DMR test.

On both groups 84 patients with a relevant stenosis were indicated by the reference standard, and in both groups 37 patients of these patients could be detected by 3DMR, resulting in identical sensitivities of 44%. Specificities were 96% in the gadodiamide group and 83% in the gadopentetate group. The differences in sensitivity and specificity with 95% CIs were 0.0 (-15.7,15.7) and 12.8 (-12.3,43.7), indicating no difference in sensitivity and a non-significant difference in specificity.

Summary of Chapter 5

There are different types of accuracy studies. We may perform single arm studies as prospective studies or as case-control studies. In a prospective study, we include consecutive patients who fulfil the inclusion and exclusion criteria. In these patients both the index test and the reference test are applied. In case-control studies we select subjects for whom we already know whether the target condition is present or absent. Then we apply the index test in these subjects. For the comparison of two diagnostic tests prospective comparative accuracy studies are appropriate. Prospective comparative accuracy studies can be conducted as paired studies or randomized studies. In paired comparative accuracy studies both tests of interest as well as the reference test are applied in all patients. In randomized comparative accuracy studies, in each patient only one of the two tests of interest (in addition to the reference test) is applied. Patients are randomly allocated to one of the index tests.

Chapter 6

Design Options for Randomized Benefit Studies

Objectives of Chapter 6

At the end of chapter 6 the reader should be able to ...

- recognize that there are various design options for conducting benefit studies
- understand that benefit studies are typically performed as randomized studies
- differentiate between randomized diagnostic studies, interaction studies and preselection designs

The general idea to study the benefit of diagnostic procedures in randomized designs can be implemented in different ways. Actually many different randomized designs have been already suggested in the literature. These designs can be divided into three main groups differing in their basic structure and also reflecting different aims (cf. Chapter 3): *Randomized Diagnostic Studies, Interaction Studies and Preselection Design Studies*. Roughly speaking, the three groups can be characterized in the following way:

Randomized Diagnostic Studies:

- Two diagnostic procedures are compared
- Randomization of all patients to one of the two procedures
- Aiming to investigate which procedure is better

Interaction Studies:

- One diagnostic test is applied in all patients
- Randomization of all patients to one of two treatments
- Aiming to investigate whether the test can predict the treatment difference

Preselection Design Studies:

- One diagnostic test is applied in all patients
- Randomization of a subset of patients identified by the diagnostic test to one of two treatments
- Aiming to study which treatment is better in the selected subset

In the following we present these three main groups in more detail, and in particular we discuss variants of the general idea which have been proposed in the literature.

6.1 Randomized Diagnostic Study

Key property
All patients are randomized either to a new diagnostic procedure or to a standard diagnostic procedure (figure 6.1).

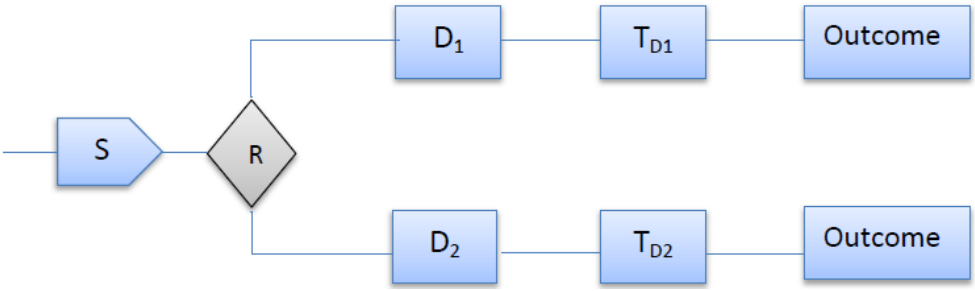


Figure 6.1: Randomized diagnostic study – the general schema.

Research question

The choice between different treatments, the planning of a treatment or in general a management decision depends on diagnostic information which can be obtained by two different diagnostic procedures. We would like to know, which procedure provides the better information, i.e., for which procedure we can expect better treatment or management decisions resulting in improved patients' health, if we base our treatment and management decisions on one of the two procedures.

Description

All subjects are randomized to D_1 or D_2 and then undergo treatment in dependence on the results of D_1 or D_2. D_1 and D_2 may be two specific diagnostic tests or they may represent more complex diagnostic procedures. The treatment choices may be left open to the clinicians or there may be rules specified how the treatment should be chosen in dependence on the results of the diagnostic procedure. Figure 6.1 shows the general structure of a randomized diagnostic study. In the case that strict rules are given how to choose the treatment in dependence on the results of a test, we may represent this design as shown in Figure 6.2.

Analytical strategy

Comparison of the outcome between the two arms.

Other names and references

Studies according to the general scheme described in Figure 6.1 have been called *Classical RCT* (Lee et al., 2009). Studies according to the specific scheme in Figure 6.2 have been called *RCT comparing tests* (Lijmer and Bossuyt, 2009).(Simon, 2010)

Ungated RCT (Vach et al., 2011)

Two-arm design (Lu and Gatsonis, 2013)

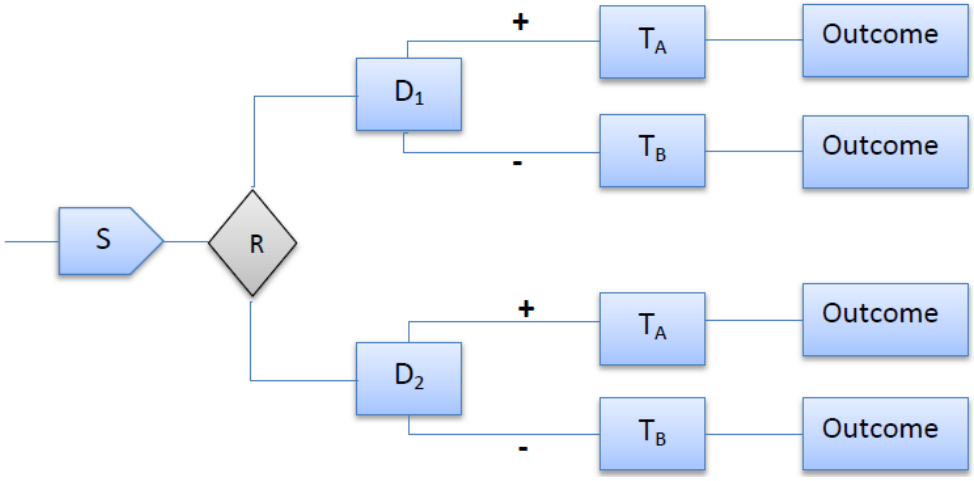


Figure 6.2: Randomized diagnostic study: The case of fixed treatment decisions in dependence on test results.

Example 1: Management of patients with chronic critical limb ischemia
The definition of critical limb ischemia (CLI) requiring vascular intervention is still under debate. According to the conventional strategy the decision for further diagnostic imaging of the arteries (primary duplex scanning and - if indicated - angiography) and thus the intention for a vascular intervention, are based on clinical symptoms, physical examination, and ankle blood pressure by the vascular surgeon involved. However, the clinical eye of the physician and ankle blood pressure measurements used so far may be insufficient to judge the severity of disease, and seems to make decision-making for a vascular intervention subjective. Previous investigations have shown that a combination of toe pressure (TP) and transcutaneous oxygen pressure ($tcPo_2$) measurements might be a good indicator for the need of an vascular intervention. Both are simple and quick procedures that provide functional information about the peripheral tissue perfusion and it is the disturbance in peripheral circulation that causes the clinical signs and symptoms. de Graaff et al. (2003) present a study including 96 patients with 128 legs clinically suspected of critical limb ischemia by a vascular specialist and referred to the vascular laboratory. All enclosed patients were randomly assigned either to the conventional strategy (62 legs of 46 patients), based on ankle blood pressure measurements and duplex scans or angiograms or to the new management strategy (66 legs of 50 patients), based on the combination of TP and $tcPo_2$ measurements and on indication duplex scans or angiograms.

For patients of the conventional strategy group the treatment choice was left open to clinicians who discussed at a weekly multidisciplinary meeting if a patient should receive a conservative treatment, an arterial bypass grafting, or a balloon angioplasty. For patients of the

new strategy group strict rules are given to choose the treatment: A vascular intervention and, subsequently, a duplex or angiography (or both) to define the type and place of intervention, are indicated only if one of the two measurements TP or $tcPo_2$ is below the cutoff level (TP$\leq$30mm Hg or $tcPo_2 \leq$35 mm Hg).

Different clinical outcomes were compared in these two strategy groups. The primary outcome was the change in pain measured by the bodily pain subscore of the SF-36. Secondary endpoints were the change in the clinical situation as judged by (limb-) survival, amputation frequency, wound healing, and change in health-related quality of life.

Pain per involved leg did not differ significantly between the two groups, although there was a tendency in favor of the new strategy. The prevalence of wounds was significantly lower in the conventional treatment group, whereas severity of wounds not significantly different. Quality of life assessed with the SF-36 physical and mental summary score was not significantly different between groups. In both groups 12 patients (25%) died during the study period. The number of major and minor amputations and interventions did not differ significantly between the two groups. No patients who received conservative treatment lost a limb to amputation because of delay in intervention. Time to first intervention was not significantly different. The study failed to show an advantage of $tcPo_2$ and TP measurements in management of suspected CLI over the clinical judgment of an experienced vascular surgeon.

As the treatment choice was left open to the clinicians in the conventional strategy group, but followed strict rules in the new strategy group, this study was intermediate between the general schemes of Figure 6.1 and Figure 6.2.

Example 2: Management of outpatients with dysphagia

Dysphagia is the medical term of swallowing difficulties. For patients with dysphagia swallowing often involves the risk, that e. g. food goes down the 'wrong way'. If then food attains the lung patients can develop aspiration pneumonia. Hence, there is a strong association between dysphagia, and the development of aspiration pneumonia. Various tests exist to evaluate and manage patients with dysphagia with the objective of reducing the incidence of pneumonia.

Aviv (2000) provides a prospective, randomized study as an initial investigation of whether flexible endoscopic evaluation of swallowing with sensory testing (FEESST) is superior to the modified barium swallow test (MBS) as the diagnostic test for evaluating and guiding the behavioral and dietary management of outpatients with dysphagia.

Depending on the day of the week 126 patients were assigned to either a strategy using MBS (on Tuesdays, Wednesdays and Fridays) or a strategy using FEESST (on Mondays and Thursdays) to guide subsequent management. 50 patients were assigned to FEESST and 76 to MBS. In both trial arms there were strict rules about how to manage and treat patients

in dependence on the test results. As the outcome variables were pneumonia incidence and pneumonia-free interval, the occurrence of pneumonia was recorded during 1 year of follow up in both arms. In the FEESST group of 50 patients, 6 developed pneumonia, resulting in a pneumonia rate of 12%. In the MBS group of 76 patients, 14 developed pneumonia, resulting in a pneumonia rate of 18.4%. The difference in pneumonia incidence between the FEESST group and the MBS group was not statistically significant. The median pneumonia-free interval in the FEESST group was 39 days, and in the MBS group was 47. The difference in median pneumonia-free interval between FEESST and MBS was not statistically significant. In summary, the study indicated an advantage for the patients when using FEST, but failed to reach significance.

Variants of randomized diagnostic studies

The general design of randomized benefit studies have been used in the literature in various contexts, which lead to different variants. In the following we discuss the four main variants:

- Comparison with nothing
- Comparison of diagnostic based treatment decision with random decision
- Random disclosure
- Gated randomized diagnostic studies

6.1.1 Comparison with nothing

Key difference

In one arm of the study no diagnostic procedure is performed. All patients are randomly assigned to a new procedure that uses the test results to determine therapy or to a control arm with standard treatment (figure 6.3).

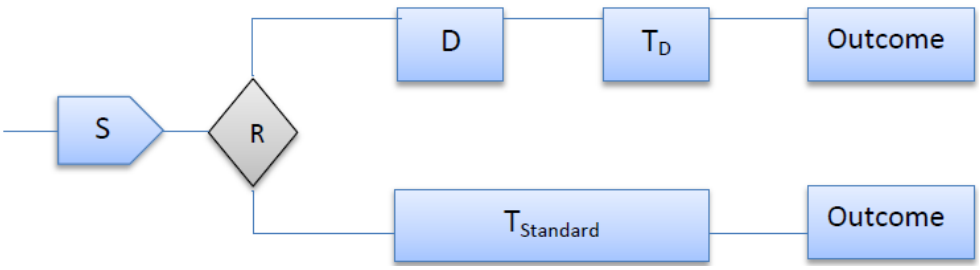


Figure 6.3: Randomized diagnostic study - comparison with nothing.

Example: Screening for prostate cancer

In the early 1990s the European Randomized Study of Screening for Prostate Cancer (ERSPC) was started to evaluate the effect of screening with prostate-specific–antigen (PSA) testing on death from prostate cancer (Schröder et al., 2009). 182,000 men in seven European countries between 50 and 74 years were included in the study and randomly allocated to a group that was offered PSA screening, or to a control group that did not receive such screening. The primary outcome was death from prostate cancer. In the screening group, 82% of men participated in screening. During a median follow-up of 9 years, the cumulative incidence of prostate cancer was 8.2% in the screening group and 4.8% in the control group. The rate ratio for death from prostate cancer in the screening group, as compared with the control group, was 0.80 (95% confidence interval [CI] 0.65 to 0.98). The absolute risk difference was 0.71 death per 1000 men. For the men who were actually screened during the first round they found a rate ratio for death from prostate cancer of 0.73 (95% CI 0.56 to 0.90). The authors concluded that PSA-based screening reduced the rate of death from prostate cancer by 20%, but was also associated with a high risk of overdiagnosis.

Motivation and consequences

A new diagnostic procedure may offer the first time the possibility to come to a new classification of the patients allowing the choice between treatment options. Then the comparator may be to apply no specific procedure, but to offer the current treatment standard or management strategy. The design can be described as shown in Figure 6.3. Note that typically there will be some diagnostics prior to randomization which is common for both arms, so we actually test the role of D as 'add on' compared to no 'add on'.

The appearance of completely new diagnostic procedures allowing us to tailor treatment more individually is typical for the research on new biomarkers. New biomarkers often promise to offer a better, alternative treatment for the patients who are marker positive and previously there was no possibility to identify these patients, and hence the comparison with 'nothing' is natural. Moreover, we have a clear idea that the results of the diagnostic procedure determine the choice of the treatment: The marker positive patients should get the alternative. So this leads to the specific case as shown in Figure 6.4.

Note that this design evaluates the diagnostic procedure in combination with the treatment. Hence it may be possible to improve patient outcomes by giving the new treatment to all patients, i.e., without using the diagnostic test.

Other names and references

The design shown in Figure 6.3 has been called *Test RCT* (Lijmer and Bossuyt, 2009). Due to

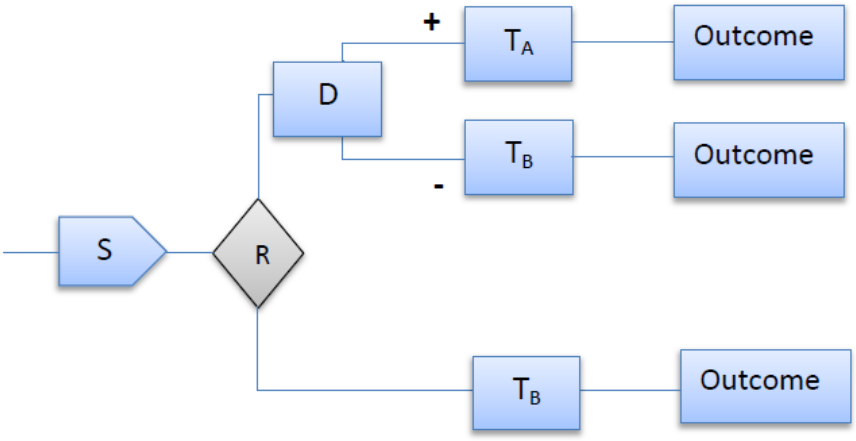


Figure 6.4: Randomized diagnostic study: comparison with nothing in the case of fixed treatment decision.

its prominent role in research on biomarkers, the design shown in Figure 6.4 has been discussed by many authors with slightly different names:
Biomarker-strategy design (Freidlin et al., 2010)
Marker-based strategy design (Young et al., 2010; Sargent et al., 2005; Eng, 2014)
Biomarker-strategy design with standard control (Buyse et al., 2011)
Marker strategy design (Simon, 2010)

Example for a study according to Figure 6.3: Management of patients with mild head injury

To manage patients with mild head injury just by observation in the hospital in the emergency departments is often standard practice. Some studies indicate that use of computed tomography (CT) is an alternative reducing costs. Furthermore, early diagnosis followed by rapid treatment is an additional potential advantage.

Geijerstam et al. (2006) and Norlund et al. (2006) presented a study to test the hypothesis whether a management strategy based on CT and early discharge is not worse with respect to clinical outcomes and is less expensive than a strategy based on observation in hospital.

2602 patients with mild head injury presenting at the emergency department were randomly allocated to one of two strategies. In the first strategy all patients with mild head injury (n=1316) received head CT. In case of a negative scan, patients were discharged home. In case of a positive scan, treatment depended on the findings. In the second strategy, all patients were admitted for observation according to local standard practice guidelines (n=1286).

At three month of follow up the prevalence of patients not fully recovered was slightly lower in the CT group (21,4%) than in the observation group (24,2%). For two patients who died in the CT group (0,2%) and for one patient who died in the observation group (0,1%) there was a possible connection of their death and the head injury. Four patients in the CT group (0,3%) and seven patients in the observation group (0.5%) developed non-fatal complications. Hence, the results of the study showed that the computed tomography strategy is not inferior to observation as regards patients' outcomes. Furthermore, it could be shown that patients with mild head injury attending an emergency department can be managed more cost effectively with computed tomography (costs of 718 euros after three months) rather than admission for observation in hospital (costs of 914 euros after three months).

Example for an adaptive study design according to Figure 6.3:
Wason et al. (2014) proposed an adaptive biomarker design where in the first stage of a two-stage scheme the design from Figure 6.4 was used, with a standard biomarker (that guided the treatment) and a cheaper alternative biomarker (that was measured and compared with the standard biomarker). After an interim analysis, in a second stage, the same design was used again, but now with the new biomarker if it was sufficiently similar (i.e., non-inferior) to the standard biomarker.

Example for a study according to Figure 6.4: Management of patients with non-small-cell lung cancer
Cobo et al. (2007) presented a study where DNA excision repair protein (ERCC1) overexpression in tumor RNA was used to adapt chemotherapy in patients with advanced non-small-cell lung cancer. It has been shown in various publications that ERCC1 expression influences ERCC1-mediated deoxyribonucleic acid (DNA) adduct repair activity, and agents that affect ERCC1 in tumors may result in increased or decreased sensitivity to cisplatin.

Patients were randomly assigned to either the control arm (N=114) or the genotypic arm in which ERCC1 was assessed (N=228). Patients in the control arm received a standard treatment of docetaxel plus cisplatin. In the genotypic arm, patients with low ERCC1 levels received docetaxel plus cisplatin, and those with high levels received docetaxel plus gemcitabine. The primary end point was the overall objective response rate. Objective response was attained by 53 patients (39.3%) in the control arm and 107 patients (50.7%) in the genotypic arm (P = .02). Hence an advantage of genotyping could be proved. Whether this implies also a clinical benefit depends on whether we have external evidence that improved response rates are associated with patient relevant outcomes like survival.

6.1.2 Comparison of diagnostic based treatment decision with random decision

Key difference

In the control arm no diagnostic procedure is applied, but patients in the control arm are randomly allocated to one of two treatment options being used in the test-based arm.

Motivation and consequences

In some situations, the additional value of a diagnostic modality has to be investigated, but it may be unclear, how usual care can be defined, as there are two concurring treatment options A and B, which both are used in practice. So we are in need to compare the use of the diagnostic modality to choose treatment A or B with both applying A directly or applying B directly. To allow a fair comparison, this requires to randomize patients not undergoing the new diagnostic procedure to either A or B. This design is visualized in Figure 6.5. Actually, this is a three

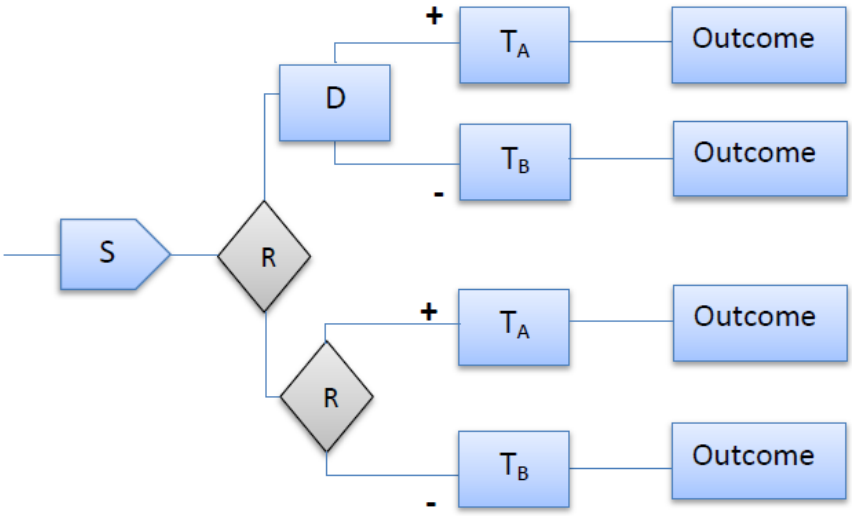


Figure 6.5: Comparison of diagnostic based treatment decision with random decision rule

arm study, allowing to compare the benefit from using the diagnostic modality with deciding for A always and with deciding for B always. So we can aim in demonstrating that using the diagnostic modality to choose between A and B is in any case better, independent of whether we regard A or B as the current standard. However, when using this design we cannot exclude that the current - even highly unstandardized - practice to choose between A and B may be better than using the diagnostic procedure to make this choice.

Other names and references

Biomarker research is a typical area for such a design, if there is current standard and a new treatment option, and if it is unclear, whether the new treatment option is beneficial only for marker positive patients. Hence this design has also been discussed and named differently in the literature on biomarker research:
Biomarker-strategy design with randomized control (Buyse et al., 2011)
Biomarker-strategy design (Kunz et al., 2017, unpublished)
Marker-based strategy design II (Young et al., 2010)
Modified marker-based strategy design (Sargent et al., 2005; Eng, 2014)

We note that this design is structurally equivalent to a design that is known from a quite different context, namely the two-stage trial design or doubly-randomized preference design (Rücker, 1989; MacLehose et al., 2000). It was applied in studies investigating patients' treatment preference (Clark et al., 2008; McCaffery et al., 2011) and also in studies comparing experimental allocation and choice in the social and behavioral sciences (Shadish et al., 2008; Long et al., 2008; Pohl et al., 2009). In the doubly-randomized preference design, the decision for treatment A or B is guided by the patient's preference, instead of a diagnostic test or biomarker.

6.1.3 Random Disclosure

Key difference

One diagnostic procedure is applied in all patients. The randomization occurs after the test has been applied, however the test results are not yet known. Only in one arm the test results are communicated, in the other they are not revealed. Figure 6.6 and 6.7 illustrate this variant.

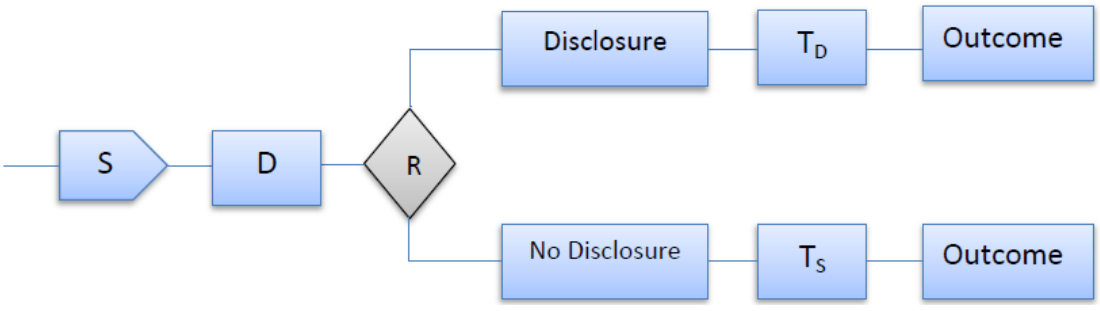


Figure 6.6: Random disclosure design.

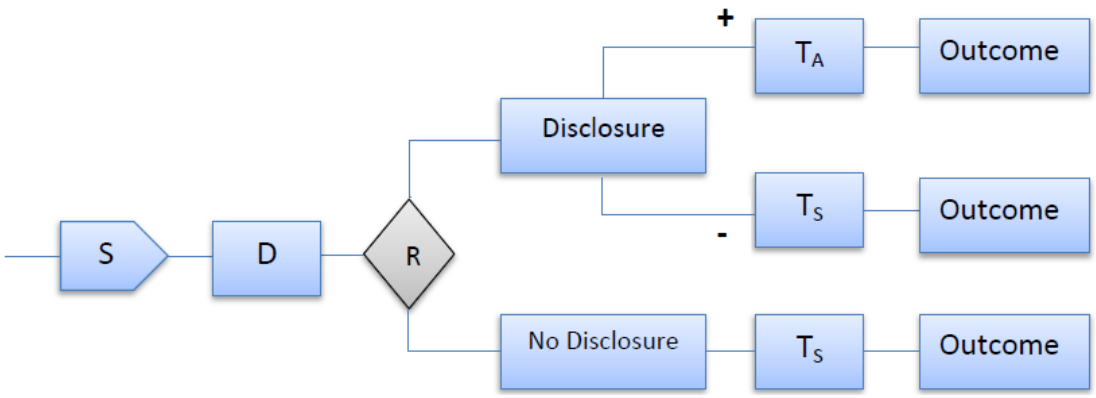


Figure 6.7: Random disclosure design in the case of a fixed treatment decision rule.

Motivation and consequences

This design is used when the diagnostic procedure to be investigated is already used as part of the standard care and/or if there is an interest in studying also the pure prognostic value of the diagnostic test result under the standard treatment.

The random disclosure design does not offer any advantage over the 'comparison with nothing' design if we are interested in demonstrating the additional benefit of the new procedure. It may be even dangerous as the treating clinician may be interested in a disclosure in the 'no disclosure' arm to optimize treatment. The advantage is mainly the possibility to study also the association of the result of the diagnostic procedure on the prognosis of the patients, if the standard treatment is given.

Other names and references

This design was named *Random disclosure* by Lijmer and Bossuyt (2009).

Example: Management of women with intrauterine growth retardation

Intrauterine growth retardation (IUGR) is diagnosed if a baby appears smaller than expected in the mother's womb during pregnancy and their weight is below the 10th percentile for their gestational age.

Nienhuis et al. (1997) reported a study investigating the effect of using Doppler ultra sound (US) in the management of women with IUGR. In this study, 150 pregnant women with IUGR underwent Doppler US and were subsequently randomized to either an intervention group (n=74) or a control group (n=76). In the intervention group the results of the Doppler US were revealed and women were requested to be hospitalized in the case of abnormal flow and to be discharged with outpatient management otherwise. In the control group the results of the

Doppler US were not revealed and all women received the standard management strategy, which was hospitalization. The aim of the study was to show that the use of Doppler US allows us to identify low risk cases and hence reduce hospitalization without affecting perinatal outcomes in a negative manner. In the study a significant reduction of the duration of hospitalization could be observed in the disclosure group (median 7.5 days vs. 18.5 days, p=0.02). However, the hospitalization rate during pregnancy was actually increased (45.7% vs. 36.1%,p=0.24). No differences could be observed with respect to neonatal and postnatal hospitalization rates (53.6% vs 52.9%, and 51.4% vs 51.4%). Perinatal outcomes were better in the disclosure group, but differences did not reach significance. The authors explained the unexpected increase in the hospitalization rate partially by the fact that not all clinicians follows the intended management strategy, i.e., that some clinicians hospitalized the women in spite of indication for a low risk situation based on the Doppler US.

Note that the authors justified the use of disclosure design in the following way: 'At the time of the study, Doppler ultrasound was not routinely available for clinical management. Therefore, withholding Doppler information from patients in the control group did not constitute an ethical problem.'

6.1.4 Gated randomized diagnostic studies

Key difference

All patients initially undergo both diagnostic procedures. Randomization takes place only for patients in which the two tests results disagree. In this sense, the diagnostic procedures play the role of a 'gate'.

Motivation and consequences

In a randomized diagnostic study for many patients it would not matter whether they are randomized to one arm or to the other, as the two diagnostic procedures would give the same results. These patients do not provide any information on a difference between the procedures to be compared. However we cannot exclude them, as we cannot identify them. In gated randomized studies (Figure 6.8) we apply first both procedures to be able to identify these patients and then we restrict randomization to patients with discordant results. Consequently, all patients undergo both tests D_1 and D_2. The therapeutic intervention as a standard of care is predefined for patients in which the two tests agree but is decided by randomization for cases in which the two tests disagree. The randomization decides on whether to follow D_1 or D_2. The two arms 'following D_1' and 'following D_2' are compared. Note that within each arm the treatment varies!

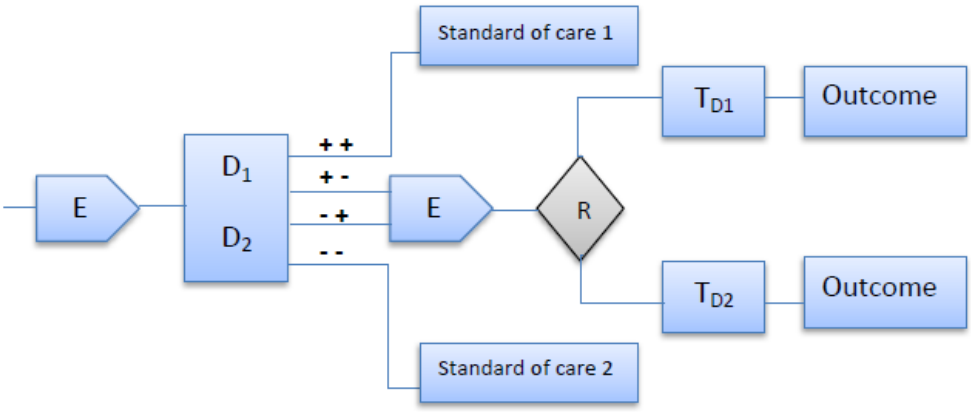


Figure 6.8: Gated randomized diagnostic study

If there are only two treatment options A and B, we may represent the design equivalently as shown in Figure 6.9, where the randomization decides actually about the treatment option. However, in the analysis we will not compare the two treatment arms, but the two groups 'treatment as suggested by D_1' and 'treatment as suggested by D_2'.

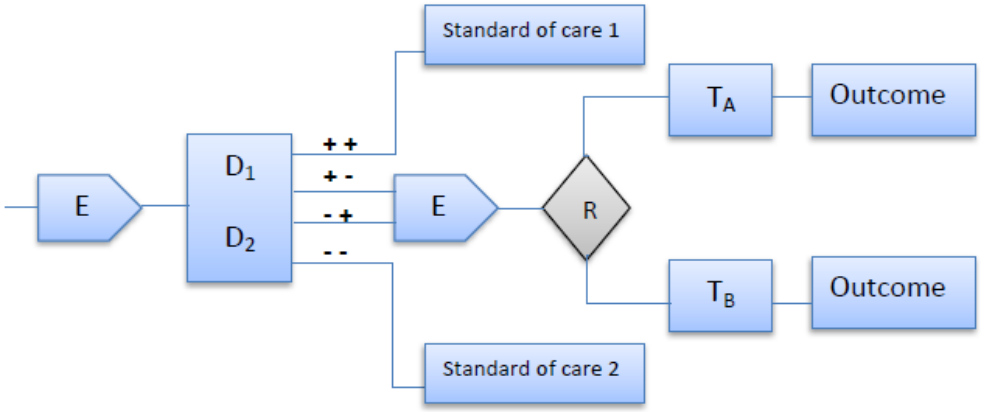


Figure 6.9: Gated randomized diagnostic study with only two treatment options A and B

Other names and references

The name 'gated RCT' was introduced by Vach et al. (2011). There are also other names for this type of study design:
Discordant risk randomization design (Buyse et al., 2011)
RCT of discordant test results (Lijmer and Bossuyt, 2009)
Paired design (Lu and Gatsonis, 2013)

Marker discordance design (Simon, 2010)

Example: The MINDACT study

The MINDACT (Microarray In Node-negative and 1 to 3 positive lymph node Disease may Avoid ChemoTherapy) trial compares the prognostic value of a 70-gene signature for breast cancer with established clinicopathological criteria to identify women with node-negative early-stage breast cancer who can avoid adjuvant chemotherapy. All women are assessed by using both the new gene signature and conventional criteria to classify their risk of disease recurrence (Figure 6.10) (Cardoso et al., 2016). Patients were categorized via the gene signature in genomic G-high risk and in G-low risk patients. Via the conventional criteria the patients were categorized in clinical C-high risk and C-low risk patients. Only women with discordant results (G-low and C-high or vice versa) are randomly assigned to receive chemotherapy or no chemotherapy; women with concordant results are not randomized but treated according to the standard of care. (Women with both C- and G-low risk are only observed, and women with both C- and G-high risk are treated with chemotherapy).

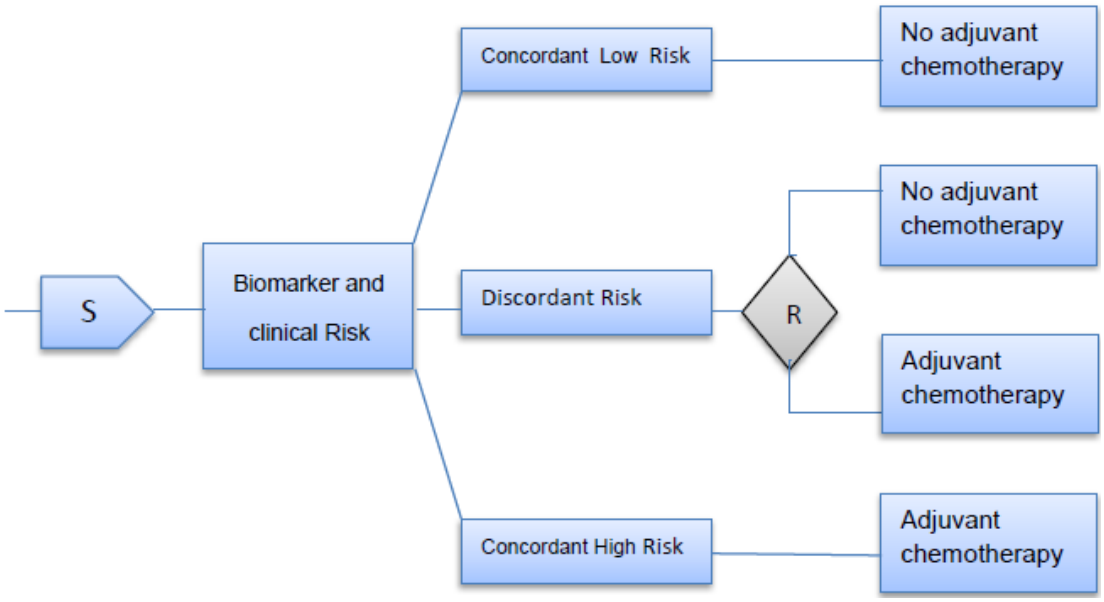


Figure 6.10: The MINDACT trial design.

The MINDACT study has enrolled about 6700 patients, and the enrollment is closed. Final results are to be expected soon, but results from the pilot phase of the study are already available. Rutgers et al. (2011) presented the results of the first 800 enrolled patients. They describe the results of the pilot study in the following words: 'Among the 800 patients, 386

(48%) were C-low/G-low, 198 (24.8%) as C-high/G-high, 75 (9.4%) as C-low/G-high and 141 (17.6%) as C-high/G-low. In total 216 (27%) cases were discordant.'

6.2 Interaction Studies

Key property

One diagnostic procedure is applied to all patients and both test positive and test negative patients are randomized to two therapy options A and B.

Research question

We have two treatment options A and B for a certain patient population, and we have hope that a diagnostic procedure can help us to decide, whether A or B is better for a single patient. So we expect that A is better than B in patients with a positive test result, but that this is not the case in patients with a negative test result (or only to a substantially lower degree).

Description

The diagnostic test is performed before the randomization. Both patients with positive and negative test results are randomized either to receive treatment A or treatment B (Figure 6.11).

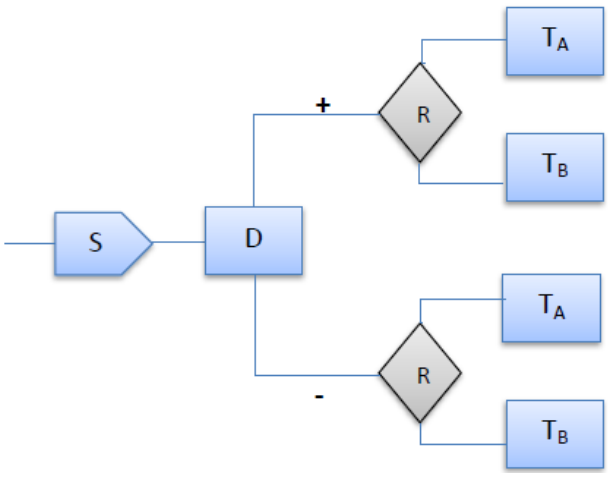


Figure 6.11: Interaction study design

In principle it does not matter whether the diagnostic procedure is performed before or after randomization to treatment, as long as it is performed prior to treatment, or as long as it is based on information in principle available prior to treatment. The design could as well

be described as a single-randomized design where the diagnostic procedure plays the role of a baseline covariate. It is often used in biomarker research, where biomarkers can be measured retrospectively based on blood or tissue samples stored for all patients in a therapeutic RCT. Then the interaction study design may look like in Figure 6.12. Note that the biomarker, though measured after randomization, is considered as a baseline variable.

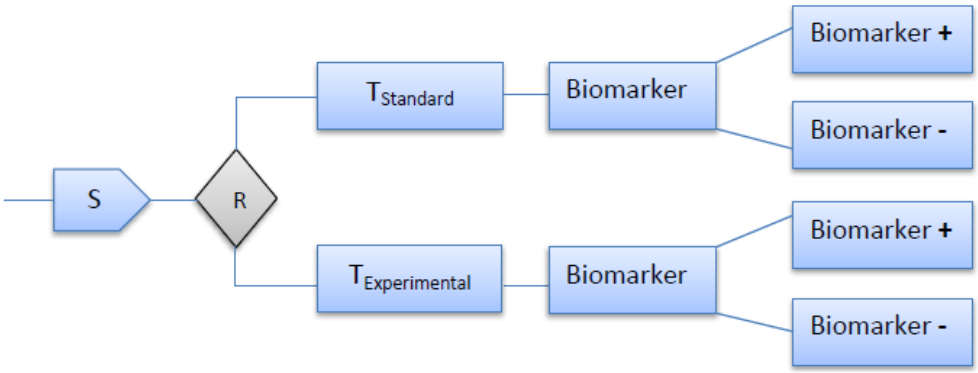


Figure 6.12: Interaction study design in the case of retrospective biomarker analysis.

Analytical strategy

This depends on the concrete aim of the study. If we only want to demonstrate that the diagnostic test provides information on the treatment effect, we have to show that the treatment effect is different between patients with a positive and a negative test result. This can be approached by considering a regression model for the outcome with the test result and the treatment as binary covariates and testing for an interaction between test results and treatment. However, if we want to conclude that the diagnostic test is beneficial for the patients, we have to do more, because even if there is a significant difference in treatment effects, it may happen that both are positive, supporting that we have to give B to all patients independent of the status of the diagnostic test. The diagnostic test is only beneficial, if we can conclude that the test allows us to come to different treatment decision in test positive and test negative patients. A sufficient condition for this is to have a significant treatment difference in test positive as well as test negative patients, favoring once A and favoring once B. If one of the two treatments reflects the current standard treatment, it would be sufficient to have evidence for a change to the new treatment in one group and to have no evidence against staying to the standard in the other group.

A possible problem with the retrospective biomarker testing is that the biomarker status may not be available for all patients, e.g., some patients may refuse agreement or tissue may

no longer be available. In such case it is important to confirm that the subgroup of patients in whom the biomarker status is known is reasonably representative of the total population randomized.

Other names and references

Interaction designs referring to Figure 6.11 have been discussed by many authors with slightly different names:
Non-targeted RCT (Lee et al., 2009)
Interaction or biomarker-stratified design (Buyse et al., 2011; Ziegler et al., 2012)
Marker by treatment interaction design (Young et al., 2010; Mandrekar and Sargent, 2009; Sargent et al., 2005)
Biomarker-stratified design (Freidlin et al., 2010)

Interaction designs in the case of retrospective biomarker analysis referring to Figure 6.12 have also been discussed by many authors with slightly different names:
Biomarker analysis with existing RCT (Lee et al., 2009)
Randomize-all design (Buyse et al., 2011)
Marker-interaction design (Eng, 2014)

Example for a prospective interaction design: The MARVEL trial

The MARVEL trial is an ongoing study; no publications are available at the moment. The National Cancer Institute provides at the web page `http://www.cancer.gov/newscenter/newsfromnci/2008/marvelrelease` (Information obtained on April 18, 2014) the following information: 'Approximately 1,200 lung cancer patients will be tested for the status of this biomarker, and then will be randomly assigned to treatment based on the test results. Both EGFR-positive and EGFR-negative patients will receive either the chemotherapy drugs erlotinib ... or pemetrexed ... after they have received their initial, standard chemotherapy. Erlotinib specifically targets EGFR, whereas pemetrexed blocks tumor cell growth by another mechanism.

It is hypothesized that erlotinib will be superior in the patients with EGFR-positive lung cancer, whereas pemetrexed would be favored in patients with EGFR-negative lung cancer, based on knowledge from earlier, smaller studies. MARVEL will incorporate genetic studies for erlotinib and pemetrexed that will be important to further identify patients with different sensitivity and toxicity profiles to these therapies.'

`ClinicalTrials.gov` provides at the web page `http://www.clinicaltrials.gov/ct2/show/NCT00738881?id=N0723&rank=1#desc` (Information obtained on April 18, 2014) fur-

ther information. In the description of the primary outcome measures, we find the following statement: 'Estimated using the method of Kaplan-Meier survival curves and a 1-sided stratified log rank test [accounting for all the stratification factors except FISH status and cooperative group] will be used to compare PFS between the erlotinib and pemetrexed arms within the FISH(+) and FISH(-) subgroups.'

From these information we can conclude that the study is actually using an interaction design.

Example for a retrospective interaction design: The CRYSTAL Trial

Cutsem et al. (2011) reported results from the CRYSTAL trail. In this trial patients with advanced colorectal tumors were randomized to chemotherapy with or without cetuximab. Based on tissue samples tumors of patients were categorized in KRAS wild-type tumors (WT) and KRAS mutant tumors (MT). As results of previous studies showed that the mutation status of the KRAS gene effects the response to cetuximab (Karapetis et al., 2008), only for patients whose tumors were wild-type for KRAS an increase of response was expected when cetuximab was added to the chemotherapy. In the CRYSTAL trial 599 patients were randomly assigned to chemotherapy consisting of irinotecan, fluorouracil, and leucovorin (FOLFIRI) plus cetuximab and 599 patients to FOLFIRI alone. DNA samples were taken before randomization. Cutsem et al. (2011) describe the results of the CRYSTAL trial by the following words: 'Patients whose tumors were wild-type for KRAS who received cetuximab plus FOLFIRI had a significantly reduced risk of disease progression (median PFS, 9.9 v 8.4 months; HR, 0.696; P<.0012) significantly improved overall survival (median survival, 23.5 v 20.0 months; HR, 0.796; P < .0093) and significantly increased odds of response (best overall response rate 57.3% v 39.7%; odds ratio, 2.069; P<.001) compared with those who received FOLFIRI alone. [...] In patients whose tumors carried mutations in KRAS, there was no evidence of a benefit associated with the addition of cetuximab to FOLFIRI in relation to PFS, overall survival, or best overall response.' With respect to the source of the information on the KRAS status, the following information was provided: 'DNA was extracted from formalin-fixed paraffin-embedded (FFPE) tumortissue and the mutation status of codons 12 and 13 of the KRAS gene assessed using a polymerase chain reaction clamping and melting curve technique ...' So this study was analysed as an interaction study with retrospective application of the diagnostic test.

6.3 Preselection Design

Key property

One diagnostic procedure is evaluated in all patients, but random assignment to treatment is restricted to patients with specific test values.

Research question

In patients characterized by the positive (or negative) result of a certain diagnostic procedure, there are two therapy or management options, and we want to know which is better. A typical example is the decision about adding a targeted component to the current standard therapy in patients who have a positive biomarker status.

Description

This design is based on the assumption that not all patients will benefit from the treatment under consideration, but rather that the benefit will be restricted to a subgroup of patients who express (or not express) a specific feature, which we can assess by a diagnostic test. This design involves testing of all patients of interest and selecting only patients with a positive test result for randomization (Figure 6.13).

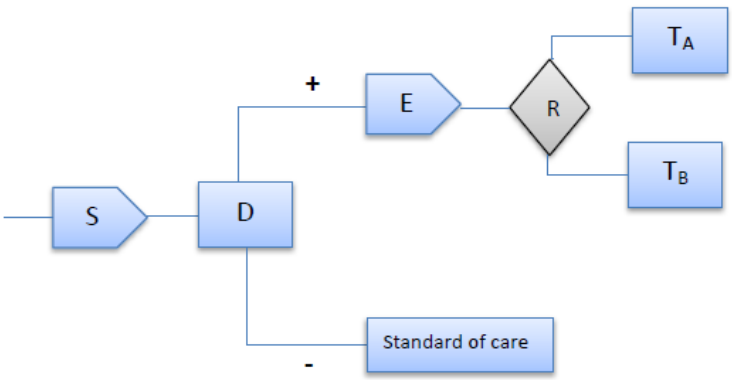


Figure 6.13: Preselection study design

Analytical strategy

Comparison of outcomes between the two treatment subgroups.

Preselection designs do not allow obtaining evidence for a clinical benefit from using the diagnostic procedure in a strict sense. If the trial demonstrates that A is better than B in test positive patients, it may still be that A is also better than B in all patients, or that A is at least not worse than B in the test negative patients. Hence it may be possible that we can improve

patient outcomes by giving A to all patients, i.e., without using the diagnostic test. However, if B reflects the current standard therapy, a change to another therapy would not be allowed without evidence for benefit (or at least non-inferiority). Hence to achieve any improvement in the whole patient group, we need to apply D to all patients to select those for whom we know that there is a benefit on average. So preselection designs can be helpful to establish the use of a diagnostic test in a patient population, given the current knowledge. However, later studies may show that A is beneficial also for test negative patients, and then there is no longer a need for using the test.

Other names and references

The most popular name for this type of design is '*enrichment design*'.
There are also other names for the preselection design:
RCT of test positive (Lijmer and Bossuyt, 2009)
Targeted RCT (Lee et al., 2009)
Targeted or selection design (Ziegler et al., 2012; Buyse et al., 2011)
Targeted or enrichment design (Mandrekar and Sargent, 2009)
Targeted enrichment design (Simon, 2010)

Example 1: The CALGB-10603 Trial

The purpose of the CALBG-10603 study is to compare the effects of a standard chemotherapy regimen for acute myeloid leukemia patients (AML) that includes the drugs daunorubicin and cytarabine combined with or without midostaurin. The prognosis for patients with AML is variable and dependent on the presence of mutations in the FLT3 tyrosine kinase. Internal tandem duplications (ITD) in the juxtamembrane region and mutations in the tyrosine kinase domain (TKD) cause constitutive activation of FLT3 and lead to blast proliferation. Preclinical data suggested that chemotherapy combined with FLT3 inhibitors, including midostaurin, could synergistically kill leukemic cells. Midostaurin may also help daunorubicin and cytarabine work better by making cancer cells more sensitive to the drugs. Midostaurin also may stop the growth of cancer cells by blocking some of the enzymes needed for cell growth.

This trial uses a biomarker to restrict eligibility to AML patients who have a documented FLT3 mutation (leading to constitutive activation of FLT3 kinase) and then randomly assigns patients to a standard treatment (daunorubicin and cytarabine) or a standard treatment plus the FLT3 kinase inhibitor midostaurin. Patients without the FLT3 mutation are considered as off-study. This study is still ongoing, but no longer recruiting patients. Information can be found at the web page of the National Cancer Institute `http://www.cancer.`

gov/clinicaltrials/search/view?cdrid=590404&version=healthprofessional or at the web page of ClinicalTrials.gov http://clinicaltrials.gov/show/NCT00651261.

Example 2: The TAILORx Trial

Lee et al. (2009) summarize the TAILORx trial in the following way: 'Oncotype DX is a 21-gene prognostic assay developed to classify women with node-negative, ER-positive breast cancer into three categories according to their risk of developing recurrent disease (low, intermediate and high risk). It has been proposed to guide treatment decisions by sparing women who are at low risk unnecessary chemotherapy, and identifying those who are at high risk and need treatment. Oncotype DX is currently being prospectively assessed in the TAILORx trial (Trial Assigning Individualized Options for Treatment [Rx]).The primary objective of the trial is to investigate the efficacy of chemotherapy as an addition to hormone therapy in women who are at intermediate risk (recurrence score, 11–25). The working premise is that patients in the intermediate-risk group will do no worse with hormone therapy alone than they would with hormone therapy plus chemotherapy. This study assumes that chemotherapy does not improve outcomes in patients at low risk (recurrence score, < 11) but will be beneficial in patients at high risk (recurrence score, > 25).'

The design of the TAILORx trail can be summarized as shown in Figure 6.14.

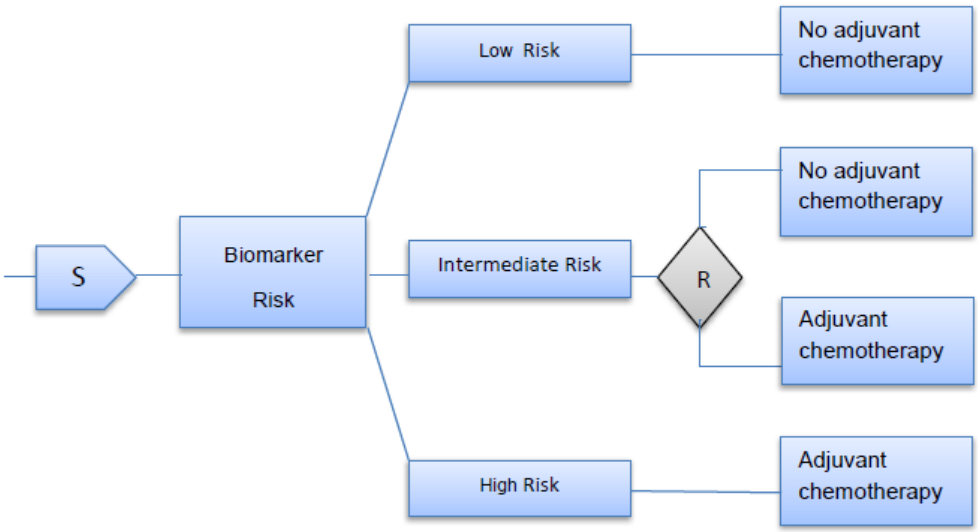


Figure 6.14: The design of the TAILORx study.

Summary of Chapter 6

The common property of benefit studies is that we want to assess the actual benefit for the patients by using patient relevant outcomes like survival or quality of life. Benefit studies are typically performed as randomized studies, but there are differences with respect to who is randomized to what. We can distinguish three types of benefit studies: In randomized diagnostic studies, patients are randomized to two diagnostic procedures, and then they are followed to assess their outcome. In interaction studies, only one test is applied, and all patients are randomized to two possible treatments. In preselection designs one test is applied in all patients, and only the test positive patients are randomized.

Chapter 7

Linking Accuracy to Benefit

Objectives of Chapter 7

At the end of chapter 7 the reader should be able to ...

- recognize that there exists a relation between accuracy and benefit
- recognize that the expected overall benefit can be expressed as a weighted average of the change in sensitivity and in specificity
- understand that concluding from improved accuracy to a benefit can be controversial
- recognize that it is possible to incorporate considerations about the expected benefit into the analysis of a comparative accuracy study

So far we have considered accuracy and benefit as different and alternative concepts. However, there is of course a relation between the concepts. The obvious one is that without an improvement in accuracy we typically cannot expect a (long term) benefit for patients. The less obvious and much more controversial relation is the question, whether, how and when we may conclude a benefit for patients, if (only) an improvement in accuracy has been demonstrated. In this chapter we will try to shed light on this question. We start in Section 7.1. with establishing a formal link between accuracy and benefit by developing a mathematical formula relating the benefit we can expect to observe in a randomised benefit study to the results of an accuracy study performed in the same population. In section 7.2. we discuss the idea of 'linked evidence', which means the attempt to predict the expected benefit from the results of an accuracy study. Finally, we discuss in Section 7.3 why incorporating the idea to link benefit to accuracy already in the analysis of accuracy studies is useful.

Note that in this chapter we discuss only the situation of benefit studies in which two single diagnostic tests are compared and in which the results of the tests determine the subsequent management. It is typically much harder to link the results of benefit studies involving more complex diagnostic procedures to the results of accuracy studies.

7.1 A Formal Link Between Accuracy and Benefit

In this section, we explain an approach by Gerke et al. (2015). Let us start with assuming that we want to compare a new diagnostic test with the current standard test, and that we have identified the target situation, the target population, and a study population which is sufficient close to the target population. In this study population we can conduct a paired accuracy study, or we can perform a randomized benefit study, randomizing all patients either to the standard test or the new test. In this section we will investigate the relation between the results from the accuracy study and the results from the benefit study.

In the accuracy study we will for each patient obtain the results from three tests: the standard test, the new test and the reference test. So we have eight possible combinations of test results, and each combination will appear in the study with a certain probability. This situation is reflected in Table 7.1. We can also say that in the accuracy study we observe the different combinations with certain relative frequencies, and we can interpret the forth column of Table 7.1 accordingly. This difference between probabilities and relative frequencies does not matter for the following considerations. The first step in linking accuracy to benefit is to try to take the consequences of changing test results into account. We discussed already in Section 3.1 that the consequences of TP, FP, TN and FN test results are typically very different, so a

standard test	new test	reference test	probability / relative frequency
+	+	+	p_{+++}
+	+	-	p_{++-}
+	-	+	p_{+-+}
+	-	-	p_{+--}
-	+	+	p_{-++}
-	+	-	p_{-+-}
-	-	+	p_{--+}
-	-	-	p_{---}

Table 7.1: The eight possible combinations of the results from the standard test, the new test and the reference test, and their probabilities/relative frequencies in a paired accuracy study.

first step is to think about actual changes of the test results in terms of these four categories. The second step would be to take into account which actual change in management will happen if we rely on the results of the new test instead of the standard test. Here we will assume in the following that the test results completely determine the subsequent management process. If the results of the standard test and the new test coincide, then of course the categories coincide, too, and we have no reason to expect any change in the management process. This applies to the first two and the last two rows of Table 7.1, and justifies the corresponding entries in Table 7.2. In the remaining four rows the category of the test result as well as the management process will change. For example, if the result of the standard test is negative, the result of the new test is positive, and the reference test is positive, too, we move from a false negative test result to a true positive test result, if we rely on the new test instead of the standard test. The change in management process will be from the management we apply in test negative patients to the management we apply in test positive patients. We can also try to judge the quality of this change in the sense that it is a change from an 'incorrect' or inadequate management to a 'correct' or adequate management. Table 7.2 summarizes these considerations for the four combinations of test results with a change in management. Roughly speaking, the new test is better than the standard test, if correct changes appear more often than incorrect changes, i.e., if $p_{-++} + p_{+--}$ is (distinctly) larger than $p_{-+-} + p_{+-+}$. Now let us consider the situation that we would perform a randomized benefit study in this study population. We have to choose some patient relevant outcome measure. Then we can often express the difference between the two arms as a difference $\hat{\Delta}$ between the arm specific mean values (for example if we use a quality of life score as outcome), or a difference $\hat{\Delta}$ in relative frequencies (for example if cure

standard test	new test	reference test	change in test result status	change in management	quality of change in management
+	+	+	TP→TP	no	neutral
+	+	-	FP→FP	no	neutral
+	-	+	TP→FN	$T_+ \rightarrow T_-$	c→i
+	-	-	FP→TN	$T_+ \rightarrow T_-$	i→c
-	+	+	FN→TP	$T_- \rightarrow T_+$	i→c
-	+	-	TN→FP	$T_- \rightarrow T_+$	c→i
-	-	+	FN→FN	no	neutral
-	-	-	TN→TN	no	neutral

Table 7.2: The eight possible combinations of the results from the standard test, the new test and the reference test, and the change in test result status, the change in management, and the quality of the change in management. T_+ denotes the management/treatment to be applied in the case of a positive test result, T_- denotes the management/treatment to be applied in the case of a negative test result. c→i denotes a change from a correct management to an incorrect management, and i→c denotes a change from an incorrect to a correct management.

rates or five year survival probabilities are estimated in each arm). Now let us assume that we would apply in this benefit study both the standard test, the new test and the reference test in each patient, too. (This is of course rarely done in randomized benefit studies, but nevertheless the patients would have test results, if we would apply all tests. The assumption of test results in patients actually not tested is what statisticians call a 'contrafactual' assumption.) Then we would be able not only to observe the overall difference $\hat{\Delta}$, but also the difference in each of the eight subgroups defined by the possible test results, as indicated in Table 7.3. Now we can think about which difference we would actually expect in each of the eight subgroups. If the results of the new test and the standard test coincide, it cannot matter to which arm the patient is randomized, as the patient would in any case follow the same management process (cf. Table 7.2.). Consequently, we would expect no difference in the outcome. So for all four rows with no change of the test result, the expected difference between the two arms is 0, and any difference we would observe would be just by chance. If the results of the new test and the standard test differ, the patients would be exposed to different management processes in the two arms of the study, so we will expect a non-zero difference, as indicated in Table 7.3. The numbers $\Delta_{TP\rightarrow FN}$, $\Delta_{FP\rightarrow TN}$, $\Delta_{FN\rightarrow TP}$, and $\Delta_{TN\rightarrow FP}$ denote here the expected difference between the two arms comparing the new test vs. the old test. So we expect that $\Delta_{FP\rightarrow TN}$

and $\Delta_{\text{FN}\to\text{TP}}$ are positive numbers, for example indicating a gain in survival probability, as here patients move from an incorrect to a correct management. And we expect that $\Delta_{\text{TP}\to\text{FN}}$ and $\Delta_{\text{TN}\to\text{FP}}$ are negative numbers, for example indicating a loss in survival probability, as here patients move from a correct to an incorrect management. No we can relate these numbers to

standard test	new test	reference test	subgroup specific change in benefit	Expected subgroup specific change in benefit
+	+	+	$\hat{\Delta}_{\text{TP}\to\text{TP}}$	0
+	+	-	$\hat{\Delta}_{\text{FP}\to\text{FP}}$	0
+	-	+	$\hat{\Delta}_{\text{TP}\to\text{FN}}$	$\Delta_{\text{TP}\to\text{FN}}$
+	-	-	$\hat{\Delta}_{\text{FP}\to\text{TN}}$	$\Delta_{\text{FP}\to\text{TN}}$
-	+	+	$\hat{\Delta}_{\text{FN}\to\text{TP}}$	$\Delta_{\text{FN}\to\text{TP}}$
-	+	-	$\hat{\Delta}_{\text{TN}\to\text{FP}}$	$\Delta_{\text{TN}\to\text{FP}}$
-	-	+	$\hat{\Delta}_{\text{FN}\to\text{FN}}$	0
-	-	-	$\hat{\Delta}_{\text{TN}\to\text{TN}}$	0

Table 7.3: The eight possible combinations of the results from the standard test, the new test and the reference test, and the change in benefit and its expected value.

the expected overall benefit Δ we can expect in the benefit study, as this is a weighted average of the expected benefits in each of the eight subgroups, with weights equal to the size of each subgroup. Since the expected benefit is 0 in four subgroups, only the remaining subgroups make a contribution:

$$\Delta = p_{+-+}\Delta_{TP\to FN} + p_{+--}\Delta_{FP\to TN} + p_{-++}\Delta_{FN\to TP} + p_{-+-}\Delta_{TN\to FP} \tag{7.1}$$

This way we have established a formal link between the results of a (paired) accuracy study – as the probabilities p_{+-+}, p_{+--}, p_{-++} and p_{-+-} can be estimated from a paired accuracy study – and the benefit Δ we can expect to observe in a benefit study. Of course, this relation is in the moment of purely theoretical nature, as the benefits we can expect in each of the four subgroups are unknown. Nevertheless, the relation can be of practical value, as we will point out in the next two sections. We will touch this issue again in Chapter 8, Section 8.4.4.

However, one important insight from this theoretical consideration is the simple fact that the benefit we can observe in a randomized benefit study is only driven by those patients for whom the two tests do not coincide. This number is often small, as if the standard test is not completely worthless and the new test is of some value, we have to expect that they coincide

in the majority of patients. We will come back to this point later when we consider sample size calculations in Chapter 9.

We have previously claimed that the differences in sensitivity and in specificity are the main measures we are interested in when analysing a comparative accuracy study. So we might be interested in linking the expected benefit directly to these quantities. This is indeed possible, if we make a further assumption about the expected benefit, namely that the loss due to changing from a correct to an incorrect management is exactly the negative of the gain in changing from the incorrect to the correct management. Mathematically, this assumption can be expressed as

$$\Delta_{FN\to TP} = -\Delta_{TP\to FN}, \quad \text{and} \quad \Delta_{FP\to TN} = -\Delta_{TN\to FP}$$

We call this the reversibility assumption, and this assumption is often quite plausible: If we move a patient from the adequate to an inadequate management or if we move a patient from an inadequate to the adequate management, the consequences go in opposite directions. Under the reversibility assumption formula 7.1 simplifies to

$$\Delta = \Delta_{FN\to TP}(p_{-++} - p_{+-+}) + \Delta_{FP\to TN}(p_{+--} - p_{-+-}) \tag{7.2}$$

Now we can also express the expected benefit as a function of the change c_{sens} in sensitivity and the change c_{spec} in specificity we can observe in the paired accuracy study. Taking the notation of Table 7.1, the change in sensitivity is

$$c_{sens} = (p_{-++} - p_{+-+})/\text{prev} \quad \text{with} \quad \text{prev} = p_{+++} + p_{+-+} + p_{-++} + p_{--+}$$

denoting the prevalence of positive results in the reference test. This follows from the fact that only patients with a positive result in the reference test contribute to the sensitivity, and among these (only) those moving from a negative result in the standard test to a positive result in the new test increase the sensitivity, and only those moving from a positive result in the standard test to a negative in the new test decrease the sensitivity. Similarly, we can find for the change in specificity

$$c_{spec} = (p_{+--} - p_{-+-})/(1 - \text{prev})$$

Combining these expressions with the formula for the expected benefit under the reversibility assumption we obtain

$$\Delta = c_{sens}\ \text{prev}\ \Delta_{FN\to TP} + c_{spec}\ (1 - \text{prev})\ \Delta_{FP\to TN}$$

Thus the expected benefit is a weighted average of the change in sensitivity and the change in specificity. The weights are proportional to the prevalence and 1 minus the prevalence,

respectively. This reflects that the expected benefit is dominated by the benefit to be expected from a change from FN to TP, if the prevalence of the target condition is high, as then this change is much more frequent then the change from FP to TN. The weights also depend on the benefit to be expected in the case of a correct change. This reflects that the increase in sensitivity or specificity, respectively, implies a better overall benefit exactly due to the two corresponding expected benefits.

Remark: Expressing the expected overall benefit as a weighted average of the change in sensitivity and in specificity demonstrates that we can make a link between accuracy and benefit also in the case of unpaired, randomized comparative accuracy studies. From such studies we can still estimate the change in sensitivity and the change in specificity. However, we cannot estimate the four probabilities p_{+-+}, p_{+--}, p_{-++}, p_{-+-} from such studies. So here we have to rely on the reversibility assumption if we want to link accuracy to benefit.

For further reading, we recommend Gerke et al. (2015).

7.2 Linked Evidence

Accuracy studies do not allow to measure benefit, only accuracy. But does this imply that we need in any case a benefit study to measure benefit? Aren't there situations in which the benefit implied by improved accuracy is so obvious that we can be sure that there is a benefit for the patients, even without a benefit study?

Let us start with a simple example to illustrate some issues when following such ideas. Suppose we have a new imaging technique to detect lung cancer. It has been compared with the current standard (for example CT) in an accuracy study, targeting the population of smokers where a GP has a suspicion about lung cancer due to recently developed frequent episodes of coughing. This (paired) comparative accuracy study may have shown that in this clinical situation the new imaging technique detects not only all patients with lung cancer who have been detected by CT, but detects also all other patients with lung cancer not detected by CT, and never produces an additional false positive result. So we reach now a sensitivity of 100%, and the specificity is exactly the same as for CT. In other words, if there is a change in the test results, it is a FN→TP change, and no other changes occur. At first sight this looks like a clear advantage: We detect more lung cancer patients, and no subject can suffer from a change from a correct to an incorrect management. However, this does still not imply in any case a benefit on average. We have still to be sure that the FN→TP changes benefit from this change. For example, all these patients may be patients already in the final stage of lung cancer, where no curative treatment is possible. Then we cannot expect an improved survival.

Palliative treatment may still offer an improvement in quality of life, but may be many of these patients would die without knowing to have cancer and without substantial loss in quality of life?

To answer this, we may take a look at the stage distribution of the patients additionally detected by the new imaging technique, and we hopefully find that there are many patients from early stages in this group, for whom a curative treatment is available. Consequently, the patients would have a chance to benefit. But this is still not sufficient to conclude that there is a benefit. Our expectation that being detected in an early stage is an advantage for a patient is based on at least two assumptions. First we assume that patients additionally detected would remain undetected until a time point when they are in a more advanced disease stage. Secondly, we assume that this implies a disadvantage for them. The latter assumption can be supported by evidence from stage specific survival rates in patients treated according to the current stage-specific standard, which may indicate that more advanced stages are associated with decreased survival probability. The first assumption is more difficult to support empirically, as it would require to have an idea about the time from an initial CT failing to detect lung cancer until the final detection, and the typical change in stage happening in this period. Electronic patient records available at a population level may allow to obtain such information. Or we may have to conduct a specific follow up study of patients with a negative CT in this specific clinical situation. However, even if we find this supporting evidence for the two assumptions, it may still be argued that we cannot be sure about a benefit. It may be, for example, that the patients additionally detected form a subgroup of lung cancer patients who are resistant against the curative treatment(s) offered today. Whether such a scenario is likely or unlikely can only be judged in the light of more detailed knowledge about the new imaging technique, in particular which (biological) properties of tumours are actually used to detect them. Empirically, it is hard to find existing evidence against such a scenario, as the accuracy study is typically the first one who detected these patients systematically. To create evidence against this scenario, we have to follow the additionally detected patients in the accuracy study for some time to demonstrate that their stage-specific survival is comparable to our expectations about stage-specific survival.

This (simple!) example demonstrates two important issues in any attempt to deduce a patient benefit from improved accuracy. First, we have to use additional evidence from other sources to be combined with the results on accuracy. This evidence may come from additional results of the accuracy study or from other studies. In our example case the information about the stage distribution of the new detected cases can be obtained from the accuracy study, and information on stage-specific survival from other sources. Second, such a deduction has to be seen as a debate between an advocatus diaboli and an advocatus dei. The advocatus dei has an optimistic view focusing always on the potential benefit for patients we may hope to achieve

by the improvement in accuracy. The advocatus diaboli has a pessimistic view focusing always on why such a hope may fail, and often using worst case scenarios as an argument. To some degree it may be possible to support empirically the optimistic view or at least to weaken the pessimistic view, but finally there will be often the need to argue with a somewhat qualified believe instead of having clear evidence.

Taking this insight into account, it may be not very surprising that until now there exists no widely used, constructive framework for deducing benefit from accuracy. Several suggestions have been made in the literature. The Australian Medical Services Advisory Committee (MSAC) created the term 'linked evidence' for such a deduction (MSAC, 2005), and it has been applied in more than 80 health technology assessments in Australia (Merlin et al., 2013). However, the framework itself is actually rather informal, mainly suggesting to think about combining the results from accuracy studies with the results of treatment studies to mimic a randomized diagnostic study. As part of the GRADE initiative Schünemann et al. (2008a,b) presented some general thoughts about when and how it may be allowed to make conclusions about patient benefit from the results of accuracy studies. Several groups from health economics have suggested to use principles from health economics, where there exists a tradition to combine evidence from different sources to predict the impact of certain interventions on the societal level, but their suggestions are rather vague, too (Schaafsma et al., 2009; Sutton et al., 2008; Trikalinos et al., 2009).

The formal link we provided in the previous section does not solve any of the conceptual problems in deducing benefit from accuracy, but it provides at least a clear mathematical framework to identify the main challenges in such a process: We have to have an idea about the four values $\Delta_{TP\to FN}$, $\Delta_{FP\to TN}$, $\Delta_{FN\to TP}$ and $\Delta_{TN\to FP}$ in order to be able to predict the expected benefit Δ. Moreover, it may help to become clear about which of these four numbers are most crucial, as they enter the overall benefit with the weights p_{+-+}, p_{+--}, p_{-++}, and p_{-+-}. The larger any of these probabilities the higher the contribution to the expected benefit, and the more essential it is to have an idea about the magnitude of the corresponding benefit.

It is also worth mentioning that we typically do not need to know the exact values of the subgroups specific benefits, but that it may suffice to agree on some bounds. If we agree on lower bounds for $\Delta_{FP\to TN}$ and $\Delta_{FN\to TP}$, i.e., the positive numbers describing the gain due to moving to a correct test result, and lower bounds for $\Delta_{TP\to FN}$ and $\Delta_{TN\to FP}$, i.e., the negative numbers describing the loss due to moving to an incorrect test result (which means actually upper bounds for the maximal loss, if we regard loss as a positive number), then we obtain by applying formula 7.1 still a lower bound for the expected benefit, which is sufficient.

However, it must be emphasized again that the formal link defines only a framework, and the hard task is to get an idea about and to provide evidence for the subgroup specific benefit.

It is often thought that clinical trials comparing the treatment options we actually offer in dependence on the test results are helpful here. However, this idea is typically less easy to implement than one might believe at first sight. Let us consider a test distinguishing between the presence and absence of non-local metastases, which the presence as target condition. Application of radiation therapy makes no sense in the presence of non-local metastases, so here chemotherapy is the only option. In the case of local metastases only, both options may be applicable. If we now want to specify the benefit of moving from a false positive to a true negative test result, we have to address the question of the impact of moving from a chemotherapy to a radiation therapy in the case of local metastases only. Here we may have luck and find an RCT comparing these two treatment options in this patient group. If we want to specify the benefit of moving from a false negative to a true positive result, we have to address the question of moving from a radiation therapy to a chemotherapy in the case of non-local metastases. Then it is unlikely to find any RCT which compares these two treatment options in this patient group, as there is no reason to believe that these patients can benefit from radiation therapy.

Even if we are in a situation that the two treatment options suggested by the results of a diagnostic test have been compared in two separate RCTs for both disease states we would like to identify, it remains always the question whether these results are applicable. All these RCTs will be based on a large group of patients with the corresponding disease state, whereas we are interested only in the patients for whom the test result is changed into this disease state. And this group may be a specific subgroup where we may experience a different treatment effect than in all patients in this disease state.

Remark: There are situations, where even the reversibility assumption has to be questioned, and where we need the full framework according to formula 7.1 instead of formula 7.2. Such situations appear if identical test results from the two tests involved are interpreted differently. As an example, let us consider a primary diagnosing situation, where we change from a very informal assessment, e.g., by a symptom check list, which is publicly debated because of it insufficient accuracy, to a new imaging test, which is publicly debated as the tool of the future. TP results from any of these two tests have probably the same consequences: We start treatment. FN results from the two tests have at first sight the same consequences, too: The patient is sent home with the incorrect message: 'You are disease-free'. However, patients getting a FN result from the traditional test may not trust the results, and hence they may seek for an alternative diagnosing and will continue to be aware of new symptoms. So they have a high chance that the disease is detected soon after the incorrect diagnosis. Patients getting a FN result from the new test may trust the result to a much higher degree, and hence they do not seek additional diagnosing and may be not aware of new symptoms. So they have a low

chance that the disease is detected soon.

7.3 Integrating Benefit into a Comparative Accuracy Study

In spite of the conceptual problems presented in the previous section, it might be worth to consider linking accuracy to benefit already in the analysis of a comparative accuracy study, instead of waiting until someone else has to do this, for example as part of a health technology assessment. We can approach this in a rather simple manner: we have 'only' to agree on some values of $\Delta_{TP\to FN}$, $\Delta_{FP\to TN}$, $\Delta_{FN\to TP}$, and $\Delta_{TN\to FP}$, (or at least on some bounds as outlined in the previous section) and then we can compute an estimate for Δ using formula 7.1. Actually, we may further simplify the procedure by making use of the reversibility assumption. Then we have only to specify the two numbers $\Delta_{FP\to TN}$ and $\Delta_{FN\to TP}$, and can use formula 7.2. And even this can be further simplified, as we are typically interested in demonstrating that Δ is positive, and not necessarily in the absolute magnitude of Δ. This implies that it is sufficient to agree on the ratio

$$\frac{\Delta_{FP\to TN}}{\Delta_{FN\to TP}}$$

i.e., on the relative weight of moving from a false positive to a true negative result compared to moving from a false negative to a true positive result.

The potential value of incorporating considerations about the expected benefit into the analysis of a comparative accuracy study may be illustrated by the following example. Ng et al. (2008) investigated the clinical usefulness of 18F-FDG positron emission tomography (PET) and extended-field multi-detector computed tomography (MDCT) for the detection of distant metastases in patients with oropharyngeal or hypopharyngeal squamous cell carcinoma (SCC). A total of 160 patients with SCC underwent 18F-FDG PET and extended-field MDCT to detect distant metastases or second primary tumours. Suspected lesions were investigated by means of biopsy, clinical, or imaging follow-up, defining a composite reference standard. The results of this study are shown in Table 7.4 in a format in analogy to Table 7.1. We can observe that for 8 patients PET could detect distant metastases which remained undetected by MDCT. As the overall number of patients with distant metastases is rather small (26 out of 160), and there is only one patient where PET overlooked a patient with distant metastases detected by MDCT, we observe an increase in sensitivity from 50.0% to 76.9%, which is significant (p=0.039). On the other hand, for 6 patients without distant metastases PET gives a false positive result. Due to the large number of patients without distant metastases and the presence of 1 patient

with a move from a false positive to a true negative results, the specificity decreases only from 97.8% to 94.0%, which is not significant (p=0.125). If we want to incorporate a link to benefit

MDCT	PET	reference test	frequency
+	+	+	12
+	+	-	2
+	-	+	1
+	-	-	1
-	+	+	8
-	+	-	6
-	-	+	5
-	-	-	125

Table 7.4: The eight combinations of the results from MDCT, PET and the reference test and their observed absolute frequency in the study of Ng et al. (2008).

in the analysis, we have to think about the consequences of changing the test results. The presence of distant metastases often implicates palliative treatment, whereas the absence of distant metastases may allow for curative treatment by surgical removal of the primary tumour. A correct change to a curative treatment implies probably a higher benefit for a patient than a correct change to a palliative treatment, as in the first case we may save the life of the patient, whereas in the latter case the patient will die anyway, but has the advantage of avoiding unnecessary treatment offered with curative intent. Hence, it is reasonable to assume $\Delta_{FP\to TN}$ is larger than $\Delta_{FN\to TP}$, however, the choice of the ratio $c = \Delta_{FN\to TP}/\Delta_{FP\to TN}$ is debatable, as it requires to balance quality of life against survival. Table 7.5 shows the estimated expected benefit according to formula 7.2 (i.e., under the reversibility assumption) and corresponding 95% CIs for several choices of c less or equal to 1. For $\Delta_{FP\to TN}$ we used the value 0.2, which may correspond to a survival rate of 20% under the curative therapy. We can observe that the lower bound of the 95% CI is below 0 already for c=1, and for c less than 0.5 the point estimate of the expected benefit is negative. So, our analysis suggests that we cannot conclude a positive expected benefit from replacing MDCT by 18F-FDG PET in spite of a significant improvement in sensitivity, the reason being that (a) the expected benefit is dominated here by the change in specificity due to the low prevalence and (b) the 6 incorrect changes to palliative treatment contribute more to the expected benefit than 8 correct changes to palliative treatment if $c < 6/8$. In general, incorporating benefit in the analysis of a comparative accuracy study offers two great practical advantages. We can overcome the problem of declaring one test as the better one,

c	estimated expected benefit	95% confidence interval
1	0.0025	-0.009, 0.014
0.75	0.0003	-0.010, 0.010
0.5	-0.002	-0.011, 0.007
0.25	-0.004	-0.013, 0.003

Table 7.5: The estimated benefit with a 95% confidence interval in dependence on c for the data of the study of Ng et al. (2008).

whenever one test shows a better sensitivity and the other shows a better specificity. We have already mentioned in Section that then we have to balance the change in sensitivity against the change in specificity. Formulas 7.1 and 7.2 exactly provide such means within a clear and transparent framework. In particular, the framework clarifies that in weighting the change in sensitivity and the change in specificity we do not only have to take into account the advantage implied by a single change from an incorrect to a correct test result, but also the prevalence of the target condition. Second, in the case of a paired accuracy study, we need to know the result of the reference test only for all subjects with discordant results between the two tests. For all other subjects, we do not need to know the reference test. So if the reference test is expensive or requires an additional follow up, we can save money or/and efforts. If may even make a study easier to be accepted by an ethics committee, as we will discuss in Section 12.3.

Further reading: Further examples of challenges to determine the (subgroup specific) benefit are given in Vach et al. (2011). Links to prediction models and decision analysis are discussed by Vickers et al. (2016, 2017)

Summary of Chapter 7

Actually, there is some relation between accuracy and benefit. Without an improvement in accuracy we typically cannot expect a (long term) benefit for patients. Sometimes it is possible to conclude a benefit for patients, if (only) an improvement in accuracy has been demonstrated. In case of a comparison between a new diagnostic test with a current standard test (paired accuracy study), when linking accuracy to benefit one has to take the consequences of changing test results into account. The results of such a (paired) accuracy study must be related to the expected overall benefit determined from a randomized benefit study on the same study population.

The reversibility assumption says that the expected benefit (the gain in changing from the incorrect to the correct management) is exactly the negative of the loss due to changing from a correct to an incorrect management. If the reversibility assumption is true, it is possible to link the expected benefit directly to sensitivity and specificity.

Part C

Evaluation of Accuracy Studies

Chapter 8

Analyzing Accuracy Studies

Objectives of Chapter 8

At the end of chapter 8 the reader should be able to ...

- describe what is meant by sensitivity and specificity
- distinguish measures of test accuracy and predictive values
- calculate sensitivity, specificity and predictive values from a two-by-two table
- explain why univariate measures are not sufficient for assessing the accuracy of a test
- describe what is visualized by a ROC curve
- explain what is measured by the area under the curve

In this chapter we take a closer look at the evaluation measures of accuracy studies we already introduced in Section 2.4. We will present the different measures in a more formal way with a view to the theory behind. Before looking into the theory we will recall what we already have learned in Section 2.4. Furthermore, we give an overview about further useful measures when analyzing diagnostic accuracy studies, their interpretation and limitations.

8.1 The Concept of Conditional Probabilities

For the formal definitions of the evaluation measures typically determined in accuracy studies we need the concept of conditional probabilities.

For example, $P(T+)$ specifies the probability of having a positive test result without any restriction regarding the current study population. Typically, this notation is used when the probability of the whole current study population is of interest. But what to do if you want to determine the probability of having a positive test result only for patients who actually have the disease in the current study population? In this situation you are only interested in the probability of a positive test result in a subset of the current study population. How to compute such values?

Conditional probabilities give us the concept to determine such probabilities. The notation for conditional probability is $P(B \mid A)$, read as the probability of B given A.

Conditional probabilities are defined by the following formula:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

where $P(A \cap B)$ denotes the probability of the joint of events A and B. The conditional probability of an event B given A is the probability that event B occurs provided that event A has already occurred. Thus the conditional probability quantifies the probability for event B restricted to the subgroup with event A.

For the definition of the evaluation measures in a more formal way, we use following notations:

D+	Disease present
D–	Disease not present
T+	Positive test result
T–	Negative test result

8.2 The Most Common Evaluation Measures

8.2.1 Sensitivity and specificity

As mentioned in Section 2.4, sensitivity and specificity are the most common measures for the accuracy of a diagnostic test. Sensitivity and specificity describe the quality of the test from a global perspective. To determine these measures the index test is conducted on a population with known 'true' disease state. This 'true status' is usually obtained through the reference test. The results of the index test can be compared with the 'true' status by summarizing the results of an accuracy study in a fourfold table.

Using the concept of conditional probabilities, we may write the sensitivity as the probability of having a positive test result, given one actually has the disease, that is,

$$\text{sens} = P(T+ \mid D+).$$

This means that sensitivity makes a statement about the subgroup of patients who actually do have the disease (see Figure 8.1).

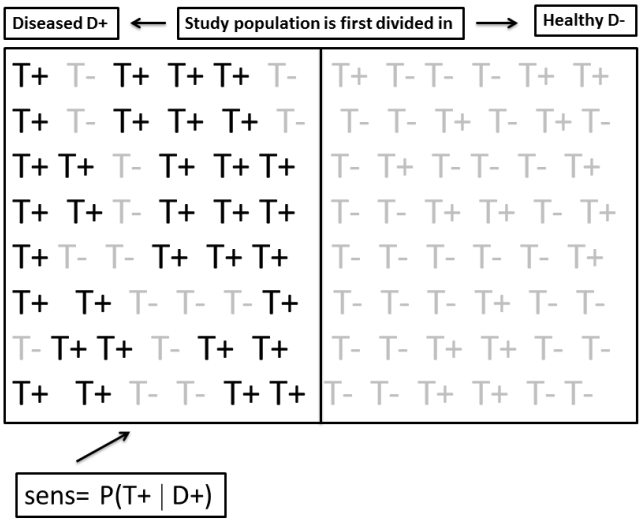


Figure 8.1: Sensitivity operates on the subgroup of patients who actually do have the disease.

Accordingly, specificity can also be defined in terms of conditional probability.

$$\text{spec} = P(T- \mid D-)$$

denotes the probability that an individual has a negative test result, given he or she actually does not have the disease. Conditioning on D- indicates that specificity makes a statement about the subgroup of patients who actually do not have the disease (see Figure 8.2).

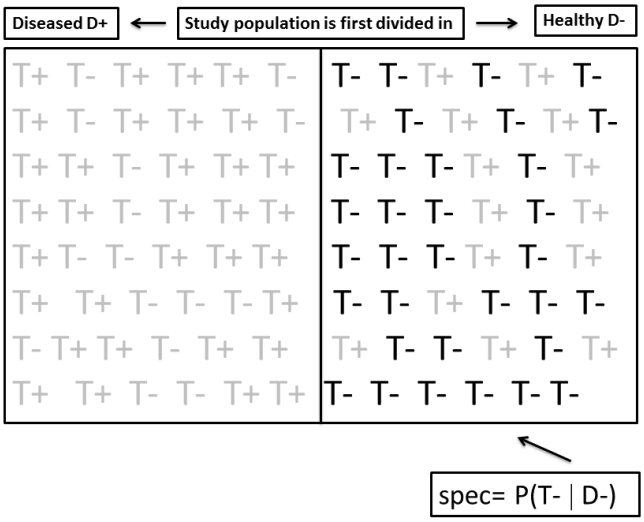


Figure 8.2: Specificity operates on the subgroup of patients who actually do not have the disease.

Comparing now sensitivity and specificity you can see that each measure operates on a different subgroup of the current study population which do not overlap (see Figure 8.1 and Figure 8.2).

Example Appendicitis results from an acute inflammation of the appendix. A diagnosis of acute appendicitis is usually made on the basis of a patient's clinical history in conjunction with physical examination and laboratory studies. A possible part of the latter is Doppler ultrasonography. In a diagnostic study the accuracy of Doppler ultrasonography on itself as diagnostic test for appendicitis should be evaluated. Therefore, 100 patients suspected of appendicitis were examined via ultrasonography. Their 'true' disease state was determined based on clinical, physical and laboratory criteria as described above used as reference. The results of ultrasonography and the reference can be summarized in a four-fold table. The fourfold table is adapted to the definition based on conditional probabilities is shown in Table 8.1: Definition, computation and interpretation of sensitivity and specificity are shown in Table 8.2.

8.2.2 Predictive values

In this section we take a closer look at the definitions of the predictive values. As already described in Section 2.4, the predictive values describe the clinical usefulness of the index test.

	Reference (based on clinical, physical and laboratory criteria) positive	Reference (based on clinical, physical and laboratory criteria) negative	
Ultrasonography (Index test) positive	44 TP	8 FP	52 I+
Event	$(T+ \cap D+)$	$(T+ \cap D-)$	$T+$
Ultrasonography (Index test) negative	10 FN	38 TN	48 I-
Event	$(T- \cap D+)$	$(T- \cap D-)$	$T-$
	54 R+	46 R-	100 N
Event	$D+$	$D-$	

Table 8.1: Summarizing the results of 100 patients suspected of appendicitis in a fourfold table. Numbers: TP = number of true positive results, FP = number of false positive results, FN = number of false negative results, TN = number of true negative results, R+ = number of positive results of the reference test, R- = number of negative results of the reference test, I+ = number of positive results of the index test, I- = number of negative results of the index test, N= overall sample size. In the gray rows, symbols in italics denote events: $D+$ = reference test positive ('diseased'), $D-$ = reference test negative ('healthy'), $T+$ = index test positive, $T-$ = index test negative.

Formal definition based on conditional event probabilities	Relative frequencies	Interpretation
sens $= P(T+ \mid D+)$ $= P(T+ \cap D+)/P(D+)$	sensitivity =TP/R+ = 44/54=0.81	81% of the patients with appendicitis have a positive ultrasonography test result
spec $= P(T- \mid D-)$ $= P(T- \cap D-)/P(D-)$	specificity =TN/R- = 38/46=0.83	83% of the patients without appendicitis have a negative ultrasonography test result

Table 8.2: Formal definition of sensitivity and specificity by conditional event probabilities, their relative frequencies and interpretation

While the positive predictive value (PPV) expresses to which degree we can trust the positive result of the index test, the negative predictive value tells us to which degree we can trust a negative result. Like sensitivity and specificity, the predictive values can be expressed in terms of conditional probabilities. The positive predictive value

$$\text{PPV} = P(D+ \mid T+)$$

denotes the probability that an individual has the disease, given he or she had a positive test result. Conditioning on $T+$ indicates that the positive predictive value makes a statement about the subgroup of patients with a positive test result (see Figure 8.3). The negative predictive

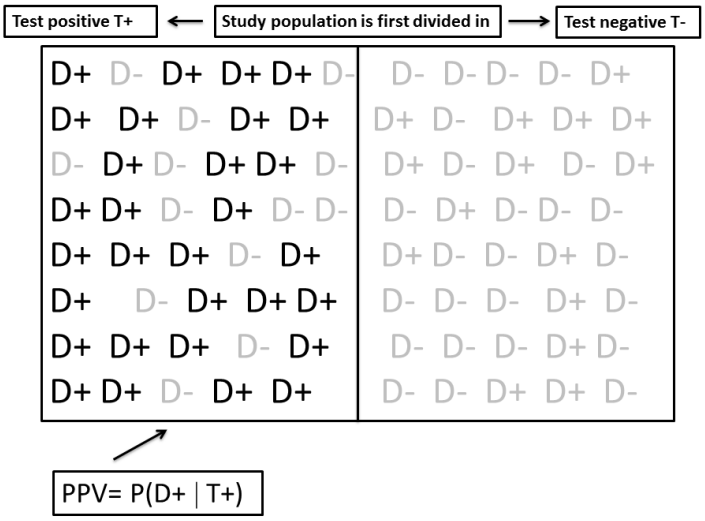


Figure 8.3: PPV operates on the subgroup of patients who have got a positive test result.

value

$$\text{NPV} = P(D- \mid T-)$$

is the probability that an individual does not have the disease, given he or she had a negative test result. Conditioning on $T-$ indicates that the negative predictive value makes a statement about the subgroup of patients with a negative test result (see Figure 8.4).

Note that the roles of the disease status (given by the reference) and the test result are reversed in the predictive values in contrast to their roles in sensitivity and specificity. How to compute the predictive values for our appendicitis example (see Table 8.1) based on the concept of conditional probabilities, their relative frequencies and their interpretation is shown in Table 8.3.

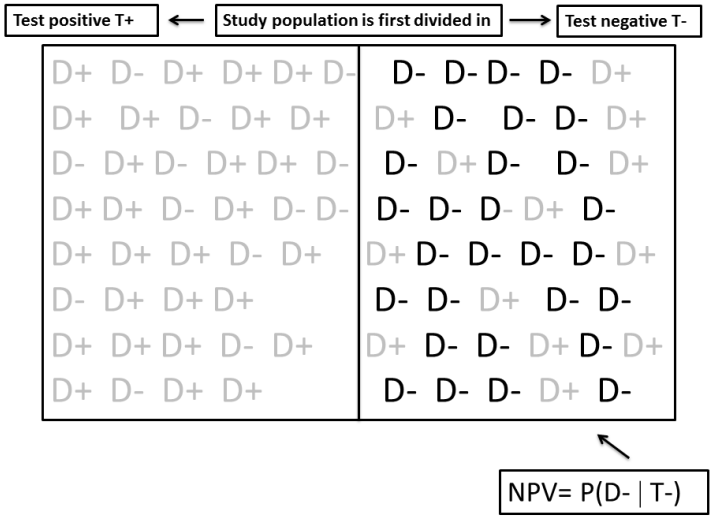


Figure 8.4: NPV operates on the subgroup of patients who have got a negative test result.

Formal definition based on conditional event probabilities	Relative frequencies	Interpretation
PPV $= P(D+ \mid T+)$ $= P(D+ \cap T+)/P(T+)$	PPV =TP/I+ =44/52=0.85	85% of the patients with positive ultrasonography test result actually do have appendicitis
NPV $= P(D- \mid T-)$ $= P(D- \cap T-)/P(T-)$	NPV =TN/I- =38/48=0.79	79% of the patients with a negative ultrasonography test result actually do not have appendicitis

Table 8.3: Formal definition of the predictive values by conditional probabilities, their relative frequencies and interpretation

8.3 The Bayes Formula

Both pairs of measures are of importance, but they describe very different aspects. Sensitivity and specificity describe the quality of the test from a global perspective. In particular, one minus sensitivity describes the proportion of patients who fail to be detected among those for whom the target condition is present. These patients are at risk of decreasing health status, reduced quality of life or of generating extra costs to the health care system, or even of dying in the case of a delayed diagnosis. One minus the specificity is the proportion of patients who are incorrectly declared to suffer from the target condition among those for whom the target condition is absent. These patients may suffer from unnecessary treatment and its adverse effects, or additional, unnecessary diagnostic tests, and from psychosocial consequences of living with an incorrect diagnosis, which may even lead to suicide. In contrast, positive and negative predictive values describe the clinical usefulness of the index test. A high positive predictive value, in particular above 0.9 or 0.99 tells us that the clinician and the patient can be pretty sure in the case of a positive test result that the target condition is indeed present and that we can hence justify to start treatment. Low positive values tell us that in the case of a positive test result we probably need to perform further diagnostic tests before we can initiate treatment. A high negative predictive value tells us that in the case of a negative test result we are pretty sure that the target condition is absent, and hence we may be allowed to send the patient home or we have to look for another explanation of the symptoms. A low negative predictive value tells us that we cannot start these steps, but that additional testing is necessary to allow this.

Predictive values play also an important role in informing patients about his or her test results. If a patient after a positive test result in a primary diagnosing situation asks the question: 'How sure is it that I have the disease?' the clinician has to answer him or her with the positive predictive value. If the test result is negative in this situation, the patient may ask 'How sure is it that I really do not suffer from this disease', the clinician has to tell him or her the negative predictive value of the test.

It is important to remember that the predictive values of a test we can observe in an accuracy study are not simple functions of the sensitivity and specificity we can observe. They also depend on the prevalence of the target condition in the study, i.e., on prev = R+/N in the notation of Figure 2.2 and Table 8.1. The relation between the predictive values and sensitivity and specificity can be expressed (by an application of the Bayes formula) as

$$\mathsf{PPV} = \frac{\mathsf{sens} \times \mathsf{prev}}{\mathsf{sens} \times \mathsf{prev} + (1 - \mathsf{spec}) \times (1 - \mathsf{prev})}$$

$$\text{NPV} = \frac{\text{spec} \times (1 - \text{prev})}{\text{spec} \times (1 - \text{prev}) + (1 - \text{sens}) \times \text{prev}}.$$

Table 8.4 illustrates for two numerical examples how the observed predictive values depend on the prevalence we can observe in an accuracy study in dependence on the observed sensitivity and specificity. The positive predictive value increases with the prevalence, and the negative predictive value decreases. This is in no way surprising, because if we have many patients with the target condition in the study population, we know already prior to applying the index test that the probability to find the target condition in a patient is high, and it becomes even higher if the test result is positive. The two formulas given above are often used to compute predictive

	sensitivity=0.9 , specificity=0.9		sensitivity=0.9, specificity=0.6	
p	PPV	NPV	PPV	NPV
0.01	0.083	0.999	0.022	0.998
0.05	0.321	0.994	0.106	0.991
0.1	0.500	0.988	0.200	0.982
0.2	0.692	0.973	0.360	0.960
0.35	0.829	0.944	0.548	0.918
0.5	0.900	0.900	0.692	0.857
0.65	0.944	0.829	0.807	0.764
0.7	0.955	0.794	0.840	0.720
0.8	0.973	0.692	0.900	0.600
0.9	0.988	0.500	0.953	0.400
0.95	0.994	0.321	0.977	0.240
0.99	0.999	0.083	0.996	0.057

Table 8.4: PPV and NPV in dependence on the prevalence for two choices of sensitivity and specificity

values for a patient population different from the study population of the actual study, assuming a certain value for the prevalence for this population. Such computations have to be taken with great care, as different subpopulations of the target population often do not only differ by the prevalence, but also by the composition with respect to the difficulty to diagnose, and hence also in accuracy, as discussed in 2.3.

Remark: Some literature about accuracy studies does not use the terms sensitivity and specificity, but the true positive rate (TPR) and the false positive rate (FPR). In the notation of Table 2.2 and Table 8.1, we have TPR = TP/R+ = sensitivity, and FPR = FP/R- = 1 – specificity.

8.4 Further Evaluation Measures

8.4.1 Likelihood ratio

Further parameters of diagnostic accuracy are the likelihood ratios (LR). The likelihood ratios address the question by which factor the test result changes the probability for a specific disease status. We distinguish between the positive likelihood ratio (LR+) and the negative likelihood ratio (LR−). Both measures are based on sensitivity and specificity and are independent of the prevalence of the disease. Based on sensitivity and specificity the LRs are defined as:

$$\text{LR}+ = \frac{\text{sens}}{1 - \text{spec}}$$

$$\text{LR}- = \frac{1 - \text{sens}}{\text{spec}}$$

Based on conditional probabilities the LRs are:

$$\text{LR}+ = \frac{P(T+ \mid D+)}{P(T+ \mid D-)}$$

$$\text{LR}- = \frac{P(T- \mid D+)}{P(T- \mid D-)}$$

Based on the numbers in the two-by-two table the LRs are estimated as:

$$\text{LR}+ = \frac{\text{TP}/(\text{TP} + \text{FN})}{\text{FP}/(\text{FP} + \text{TN})} = \frac{\text{TP}(\text{FP} + \text{TN})}{\text{FP}(\text{TP} + \text{FN})}$$

$$\text{LR}- = \frac{\text{FN}/(\text{TP} + \text{FN})}{\text{TN}/(\text{FP} + \text{TN})} = \frac{\text{FN}(\text{FP} + \text{TN})}{\text{TN}(\text{TP} + \text{FN})}$$

LR+ denotes the ratio of the probability of a positive test result among diseased and the probability of a positive test result among non-diseased. LR− describes the ratio of the probability of a negative test result among diseased and non-diseased, respectively. Both LRs range from 0 to plus infinity. An LR+ or LR− of one indicates that the particular test result is equally likely in the diseased and non-diseased population. An LR+, LR− greater than one indicates that the test result is more likely in the diseased than in the non-diseased. For a reasonable test LR+ is expected to be much greater than one, and LR− is expected to be much less than one.

How to compute the likelihood ratios for our appendicitis example (see Table 8.1) based on the concept of conditional probabilities is shown in Table 8.5.

Formal definition based on conditional event probabilities	Interpretation
LR+ $=$ sens$/(1-$ spec$)$ $= 0.81/(1-0.83) = 0.81/0.17 = 4.76$	A positive test result is 4.76 times more likely for diseased than for non-diseased.
LR$-$ $= (1-$ sens$)/$spec $= (1-0.81)/0.83 = 0.19/0.83 = 0.23$	A negative test result is 0.23 times more likely for diseased than for non-diseased.

Table 8.5: Formal definition of the LRs by conditional probabilities and their interpretation

8.4.2 Diagnostic odds ratio

The diagnostic odds ratio (DOR) of a test is the ratio of the odds of a positive test for a subject who has the disease relative to the odds of a positive test for a subject who is disease-free. What is an 'odds'? Any number (e.g., a probability or an observed relative frequency) p between 0 and 1 can be transformed into its 'odds', which is $p/(1-p)$ and lies between zero and plus infinity. For example, the odds of $p = \frac{1}{3}$ is $\frac{1}{3}/\frac{2}{3} = 1:2 = 0.5$. The DOR provides a single measure of the test performance of a diagnostic test which is independent of the prevalence. Based on conditional probabilities the DOR is defined as:

$$\begin{aligned} \text{DOR} &= \frac{P(T+ \mid D+)/\left(1-P(T+ \mid D+)\right)}{P(T+ \mid D-)/\left(1-P(T+ \mid D-)\right)} \\ &= \frac{P(T+ \mid D+)/\left(1-P(T+ \mid D+)\right)}{\left(1-P(T- \mid D-)\right)/P(T- \mid D-)} \\ &= \frac{P(T+ \mid D+) \times P(T- \mid D-)}{\left(1-P(T+ \mid D+)\right) \times \left(1-P(T- \mid D-)\right)} \end{aligned} \tag{8.1}$$

Based on sensitivity and specificity the DOR is:

$$\text{DOR} = \frac{\text{sens} \times \text{spec}}{(1-\text{sens}) \times (1-\text{spec})}$$

Based on the predictive values the DOR is:

$$\text{DOR} = \frac{\text{PPV} \times \text{NPV}}{(1-\text{PPV}) \times (1-\text{NPV})}$$

Based on the LRs the DOR is:

$$\text{DOR} = \frac{\text{LR}+}{\text{LR}-}$$

Based on the numbers in the table Rs the DOR is estimated by:

$$\text{DOR} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}}$$

The value of a DOR ranges from 0 to plus infinity, with higher values indicating better discriminatory test performance. How to compute the DOR for our appendicitis example (see Table 8.1) is shown in Table 8.6. As a limitation of the DOR, we mention that it is a univariate

Definition	Interpretation
DOR $= \frac{\text{sens} \times \text{spec}}{(1-\text{sens}) \times (1-\text{spec})}$ $= (0.81 * 0.83)/((1 - 0.81) * (1 - 0.83))$ $= 0.67/0.03 = 20.74$	A diagnostic odds ratio with a value of 20.74 greater than one indicates that the test is discriminating correctly.

Table 8.6: Definition of the diagnostic odds ratio and its interpretation

measure and thus does not distinguish between the populations of diseased and disease-free individuals. A large DOR may be explained by a high sensitivity combined with a poor specificity or vice versa. In general, it is preferred to describe the accuracy of a test using both sensitivity and specificity.

8.4.3 Youden index

The Youden index (Youden, 1950) is another univariate measure of accuracy, defined as:

$$J = \text{sens} + \text{spec} - 1$$

The Youden index ranges from -1 to 1. Values close to zero occur if a diagnostic test gives the same proportion of positive results for diseased and non-diseased, i.e the test is useless. Values close to 1 indicate that the test perfoms well and is able to distinguish between the two disease states. As the index gives equal weight to false positive and false negative values, diagnostic tests with the same value show the same proportion of total misclassified results. How to compute the Youden index for our appendicitis example (see Table 8.1) is shown in Table 8.7.

Definition	Interpretation
$J = \text{sens} + \text{spec} - 1$ $= 0.81 + 0.83 - 1 = 0.64$	A Youden index of 0.64 indicates that the test is able to distinguish between the two disease states.

Table 8.7: Definition of the Youden index and its interpretation

For further reading, see Böhning et al. (2008). For a generalization, see the next section.

8.4.4 Expected utility and expected costs

Because of the important role of costs and benefits in the decision-making process, the use of a diagnostic test is sometimes measured by expected utility. This approach is based on an analysis of the utilities (or the costs) of the four possible outcomes of a diagnostic test: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The term costs can stand for money as also for any other kind of payment associated with performing the test (e.g., healthy days lost).

It does not matter whether we consider the expected utility or the expected costs. Following Metz (1978), we express the measures in terms of costs. We give the following definitions:

C_0	basic costs associated with performing the test
C_{TP}	costs for the diseased with a positive test result
C_{TN}	costs for the non-diseased with a negative test result
C_{FP}	costs for the non-diseased with a positive test result
C_{FN}	costs for the diseased with a negative test result

The average costs resulting from performing the test can be calculated by taking the costs associated with performing the test for a participant, C_0, and adding the costs for each subgroup (TP, TN, FP, FN) weighted by the proportion (probability) of the subgroup in the whole sample:

$$\mathsf{C}_{test} = \mathsf{C}_0 + \mathsf{C}_{TP} \times P(\mathsf{TP}) + \mathsf{C}_{TN} \times P(\mathsf{TN}) + \mathsf{C}_{FP} \times P(\mathsf{FP}) + \mathsf{C}_{FN} \times P(\mathsf{FN})$$

We obtain the subgroup proportions from the probabilities for the test results for the diseased and non-diseased, times the probabilities of being or being not diseased. These in turn can be expressed by using sensitivity, specificity and prevalence as follow.

$$\begin{aligned}
\mathsf{C}_{test} &= \mathsf{C}_0 + \mathsf{C}_{TP} \times P(T+ \mid D+) \times P(D+) + \mathsf{C}_{TN} \times P(T- \mid D-) \times P(D-) \\
&\quad + \mathsf{C}_{FP} \times P(T+ \mid D-) \times P(D-) + \mathsf{C}_{FN} \times P(T- \mid D+) \times P(D+) \\
&= \mathsf{C}_0 + \mathsf{C}_{TP} \times \mathsf{sens} \times \mathsf{prev} + \mathsf{C}_{TN} \times \mathsf{spec} \times (1 - \mathsf{prev}) \\
&\quad + \mathsf{C}_{FP} \times (1 - \mathsf{spec}) \times (1 - \mathsf{prev}) + \mathsf{C}_{FN} \times (1 - \mathsf{sens}) \times \mathsf{prev}
\end{aligned}$$

After rearrangement of terms we obtain

$$\begin{aligned}
\mathsf{C}_{test} &= \mathsf{C}_0 + \mathsf{C}_{TN} \times (1 - \mathsf{prev}) + \mathsf{C}_{FN} \times \mathsf{prev} \\
&\quad - (\mathsf{C}_{FN} - \mathsf{C}_{TP}) \times \mathsf{sens} \times \mathsf{prev} \\
&\quad + (\mathsf{C}_{FP} - \mathsf{C}_{TN}) \times (1 - \mathsf{spec}) \times (1 - \mathsf{prev})
\end{aligned}$$

This equation provides some insight. We see

- The average costs of the test increase with its basic costs C_0. That is, even if a new test leads to better decisions, it may increase the total costs if its basic costs are very high, e.g., if high investment costs are necessary for an expensive device, or new specialized staff has to be recruited for a population-wide screening test.

- The average costs of the test depend on both sensitivity and specificity. As we will see in the next section, sensitivity and specificity also depend on each other, they have a negative association. This means that increasing sensitivity leads to a decrease in specificity and vice versa. If the relation is known, we may use this knowledge to minimize the costs, given the prevalence.

- In practice, the most relevant question is whether to introduce the test at all. The formula gives us the costs of not testing by inserting sens $= 0$ and spec $= 1$, as no test is equivalent to all individuals having a negative test result. We obtain for the average costs of not performing the test:

$$C_{no\ test} = C_{TN} \times (1 - \text{prev}) + C_{FN} \times \text{prev}$$

The comparison of the expected costs for performing the test with the expected costs when there is no test conducted in the same population can also help to decide about the use of a diagnostic test. The difference of the average costs without test and the average costs of the test provides the average net benefit or expected utility of the test (Metz, 1978; Baker and Kramer, 2007):

$$C_{no\ test} - C_{test} = (C_{FN} - C_{TP}) \times \text{sens} \times \text{prev} - (C_{FP} - C_{TN}) \times (1 - \text{spec}) \times (1 - \text{prev}) - C_0 \tag{8.2}$$

Looking at the last equation and assuming that the costs of false test results are always greater than the costs of true test results (i.e., $C_{FN} - C_{TP} > 0$ and $C_{FP} - C_{TN} > 0$), we see that the net benefit of introducing the test tends to be large if

- many individuals have the disease (large prevalence);

- the test has good sensitivity;

- the cost difference between false negatives and true positives ($C_{FN} - C_{TP}$) is large, that is, the consequences of not treating a diseased individual are serious;

- the test is cheap (small C_0).

The net benefit of introducing the test tends to be diminish if

- few individuals have the disease (small prevalence);
- the test has small specificity;
- the cost difference between false positives and true negatives ($\mathsf{C}_{FP} - \mathsf{C}_{TN}$) is large, that is, the consequences of treating a disease-free individual are serious;
- the test is expensive (large C_0).

If all ingredients of the last equation are known, particularly the (material or immaterial) costs or utilities are measurable, one can use the net benefit $\mathsf{C}_{no\ test} - \mathsf{C}_{test}$ for deciding whether or not to introduce a diagnostic test.

Remark: If we set $\mathsf{C}_0 = \mathsf{C}_{TP} = \mathsf{C}_{TN} = 0$ and $\mathsf{C}_{FN} = \frac{1}{\mathsf{prev}}$, $\mathsf{C}_{FP} = \frac{1}{1-\mathsf{prev}}$, we obtain for the net benefit of the test:

$$\mathsf{C}_{no\ test} - \mathsf{C}_{test} = \frac{1}{\mathsf{prev}} \times \mathsf{sens} \times \mathsf{prev} - \frac{1}{1-\mathsf{prev}} \times (1 - \mathsf{spec}) \times (1 - \mathsf{prev}) = \mathsf{sens} + \mathsf{spec} - 1$$

which is the Youden index.

Remark: Apart from the basic costs C_0, we may consider Metz' expected utility (8.2) as a special case of the expected overall benefit Δ we have derived in Section 7.1, where the standard test just means 'doing nothing'. That is, without performing the new index test all results are negative. This situation corresponds to the lower half of Table 7.3. Metz' formula (8.2) corresponds to the second half of equation (7.1), where $p_{-++} = \mathsf{sens} \times \mathsf{prev}$, $\Delta_{FN \to TP} = \mathsf{C}_{FN} - \mathsf{C}_{TP}$, $p_{-+-} = (1 - \mathsf{spec}) \times (1 - \mathsf{prev})$, $\Delta_{TN \to FP} = \mathsf{C}_{TN} - \mathsf{C}_{FP}$. Of course, it is possible to add the basic costs C_0 to the considerations in Section 7.1 in the same way as Metz did, and thus to genuinely generalize the approach by Metz (1978). For further reading, we recommend Gerke et al. (2015).

8.5 Analysis of Test Construction Studies

Many diagnostic tests are based on laboratory measurements, which typically provide a continuous marker, such as blood pressure, heart rate or temperature measurements. Often, maybe not always, high marker values indicate the disease and low values rule out the disease.

We illustrate this by an idealised representation as shown in Figure 8.5. We see two distinct probability distributions for a biomarker. The left bump shows the distribution of the biomarker among disease-free individuals, the right bump the distribution of the biomarker among diseased

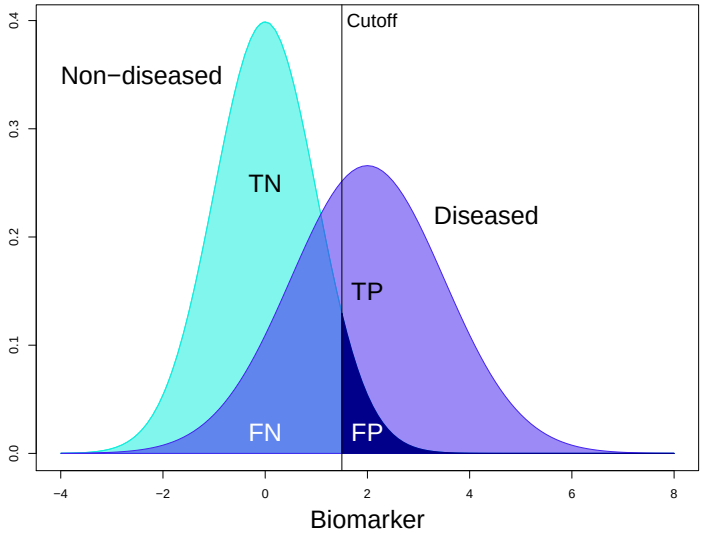


Figure 8.5: Distributions of a continuous biomarker for diseased and non-diseased individuals with a specific cut-off. The probabilities for TN, FN, FP and TP are proportional to the shaded areas.

individuals. In our example, the diagnostic marker tends to be higher for diseased patients than for non-diseased individuals.

A particular test, that is, a dichotomous decision rule is defined by specifying a cutoff (seen as a vertical line in Figure 8.5). This divides the possible measurement results in two groups. A value above the chosen cutoff is interpreted as a positive test result, a value below the cutoff is interpreted as a negative test result, or vice versa. Each such dichotomization yields a two-by-two table as given in Table 8.1. Measures of accuracy now depend on the choice of the cutoff.

As the two distributions are overlapping, a test defined in this way cannot be perfect, and the cutoff is not unambiguously defined. If, e.g., the cutoff moves to the left, the total number of positive test results and thus both TP (i.e., sensitivity) and FP increase, whereas both TN (i.e., specificity) and FN decrease. Conversely, if the cutoff moves to the right, the specificity increases, but sensitivity decreases. In other words, for a diagnostic test that is based on a continuous marker, multiple pairs of sensitivity and specificity exist and depend on each other: the greater the sensitivity, the smaller the specificity is, and vice versa.

8.5.1 ROC curve

In theory, all pairs of sensitivity and specificity can be illustrated in a so-called ROC (Receiver Operating Characteristic) curve. It shows how the true positive rate (sensitivity) depends on the false positive rate (one minus specificity).

For illustration, we look at a simple example where we build an empirical ROC curve based on sensitivity/specificity information from three observed cutoffs. We want to use body temperature as a diagnostic test for appendicitis: the higher the fever the more likely is a serious infection with appendicitis. This means that choosing a low temperature value as cutoff to classify patients many patients would obtain a positive test result, leading to high sensitivity, but low specificity. With increasing cutoff sensitivity decreases and specificity increases, as fewer patients would be assessed as positive and more patients as negative.

For example, we discuss three possible cutoffs for body temperature, 37.7° C, 38° C, and 38.5° C. Table 8.8 shows the four-fold table for cutoff 38.5° C. For the cutoff of 38.5° C and

	Appendicitis yes	Appendicitis no	
temperature measurement $\geq$ 38.5° C	54	28	82
temperature measurement $<$ 38.5° C	138	272	410
	192	300	492

Table 8.8: Two-by-two table for cutoff 38.5° C.

the other two cutoffs we obtain the values for sensitivity and specificity as shown in Table 8.9.

Cutoff	37.7° C	38° C	38.5° C
Sensitivity	0.71	0.55	0.28
Specificity	0.65	0.75	0.91
Youden Index	0.36	0.30	0.19

Table 8.9: Values for sensitivity, specificity and Youden index for the cutoffs 37.7° C, 38° C, and 38.5° C.

The ROC curve in Figure 8.6 shows the sensitivities depending on one minus specificity, including the two corners (0,0) and (1,1), which correspond to extreme values of the marker. For example, if we (nonsensically) would choose 30° C as the cutoff, the values of all individuals

would be greater and therefore all individuaals would be classified as positive (sensitivity = 1, specificity = 0). Vice versa, if we would choose 45° C as the cutoff, all individuals would have values less then that temperature and thus all would be classified as negative (sensitivity = 0, specificity = 1).

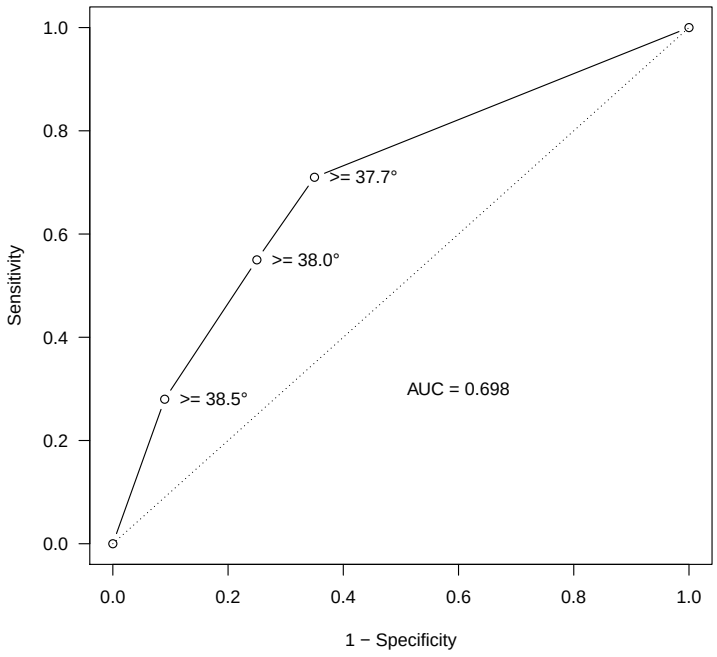


Figure 8.6: ROC curve for body temperature for diagnosis of appendicitis.

A good diagnostic test has high sensitivity and high specificity, that is, ROC ideally should fully lie in the upper-left part of the plot. For a perfect diagnostic test the true positive rate (sensitivity) is 1 and the false positive rate (1-specificity) is zero, which corresponds to the top left corner of the plot. By contrast, for a useless test (such as a coin toss) the true positive rates and the false positive rates would be similar, leading to a ROC curve on the diagonal from the bottom left corner to the top right corner (shown as a dotted line in Figure 8.6).

8.5.2 The area under the curve

The area under the curve (AUC) is a common measure to describe the overall performance of a marker as a diagnostic test for a target condition, independent of the chosen cutoff. A perfect test has AUC = 1, whereas a coin toss has AUC = 0.5. Overall test performance

can be determined by measuring the area under the ROC curve where poor tests have ROC under the curve areas close to 0.50 and excellent tests an area under the curve close to 1. For the body temperature data, we obtain an AUC of 0.698, which is moderate. We note that, though the AUC is a univariate measure and, like the DOR or the Youden index, does not distinguish between diseased and disease-free individuals, its role differs from that of other univariate measures. In contrast to those measures that depend on the choice of a cutoff, the AUC measures how strongly the underlying clinical measurement qualifies as a diagnostic marker, independently of the chosen cutoff.

8.5.3 Choice of the cutoff

If a diagnostic tests depends on the choice of a cutoff, we may ask how to make this choice. The optimal choice of the cutoff depends on the research question and hence on the purpose of the diagnostic test. If sensitivity and specificity are weighted equally, it is common to choose as cutoff the marker value $c*$ where the Youden index $J = \text{sens}(c) + \text{spec}(c) - 1$ is maximized (or, equivalently, where the sum of sensitivity and specificity is maximized). Graphically, the cutoff with this property is the point on the curve that has the maximal distance from the diagonal line. For the body temperature example, the Youden indices for the three observed points are 0.36, 0.30, and 0.19, of which 0.36, belonging to the value 37.7° C, is the maximum (it has the largest distance from the diagonal). Taking this as our criterion, we would choose 37.7° C as the optimal cutoff.

However, it is not always useful to assume equally balanced sensitivity and specificity. This balance can only be addressed in combination with the role the test plays in the diagnostic pathway – whether it is a screening test, a triage test, or a confirmation test, what are the consequences of a false negative result for patients having appendicitis or a false positive result for disease-free individuals. When thinking about the expected utility (see 8.4.4), one also would take the prevalence of the disease in the study population into account (Metz, 1978).

For a screening test that is applied as a first step in finding a diagnosis, usually a high sensitivity would be preferred, to identify as many diseased patients as possible who then undergo the following consecutive diagnostic procedures. The better the sensitivity, the greater becomes the predictive value of a negative test result. This is called a 'rule-out' scenario. We then would aim at maximizing the specificity under the condition that the sensitivity exceeds a predefined (high) lower limit. By contrast, for a test that is subsequently applied to those individuals whose screening test result is positive (confirmation test), we might prefer a high specificity to identify the false positives. This is called a 'rule-in' scenario.

However, the argument may also go the other way round, as exemplified by the discussion on

mammography screening for breast cancer (Independent UK Panel on Breast Cancer Screening, 2012, with discussion): If the prevalence is low and the specificity moderate, the predictive value of a positive test result will be poor, which means that there are many false positives who have to undergo further diagnostics or even treatment that can be not only unpleasant, but invasive or even dangerous. If average costs or benefits for performing the test and for all possible individual combinations of true status and test result are available, methods from Sections 7.1 and 8.4.4 may be used for deciding whether to perform the diagnostic test at all and if so, to optimize the cutoff.

8.6 Quantifying the difference between two tests

Comparing two tests means comparing accuracy measures. How to do this depends on the design of the study. For example, when using the paired comparative accuracy study design (5.3), all patients are tested with both the new index test, the existing test (comparator) and the reference standard. We then have not only sensitivity, specificity and so on for both tests, but also information on the frequency of all combinations of test results and thus the correlation between the tests, as given in Table 7.1. Comparisons between tests should account for this structure and use within-subject comparisons. By contrast, in a randomized comparative accuracy study (5.4), patients are randomized to having either the new index test or the comparator (and in any case they are tested with the reference standard). In this case, the randomized groups are independent and between-subject comparisons can be performed.

Regardless of the design, we have at least two parameters to compare between the tests, sensitivity and specificity. It depends on the research question how to balance these. For both measures, there exist various metrics to compare the relative performance of one test to the performance of another test. The metrics are based on the common measures for diagnostic tests and are mainly defined as differences, ratios or odds ratios and their confidence intervals. Without going into details or giving formulas (some confidence intervals will be given in Section 8.7), we here focus on the interpretation, exemplified by the sensitivities of two tests A and B, see Table 8.6.

Effect measure	Type	Interpretation
$\text{sens}_A - \text{sens}_B$	absolute difference	A better than B if > 0
$\text{sens}_A / \text{sens}_B$	ratio	A better than B if > 1
$(\text{sens}_A / 1 - \text{sens}_A) \,/\, (\text{sens}_B / 1 - \text{sens}_B)$	odds ratio	A better than B if > 1

If, say, $\text{sens}_A - \text{sens}_B > 0$, it has to be decided whether the difference is statistically

significant and whether it is clinically relevant. The first question can be answered by an appropriate statistical test. The second decision does not depend on statistical, but on clinical considerations which have to be made in advance. For example, if the existing diagnostic test has a sensitivity of 80%, one may require a sensitivity of at least 85% for the new test to be deemed superior. That is, the lower limit of the confidence interval for the difference should be 5%. At the same time, one may wish that the specificity does not become disturbingly worse (for example, the lower limit of the confidence interval for the difference of specificities should not fall below -2.5%).

Sometimes a decision may be easy if one test clearly outperforms the other one in both dimensions of accuracy, e.g. $\text{sens}_A > \text{sens}_B$ and $\text{spec}_A > \text{spec}_B$. It becomes more complicated if one test is more accurate in one dimension and less accurate in the other dimension. Tests based on an ordinal or continuous parameter may also be compared using the AUC, see Section 8.5.2. However, depending on the research question, it has to be asked whether the AUC appropriately balances sensitivity and specificity. This also holds for other univariate measures, like the DOR or the Youden index. More generally, if information on costs and/or utilities are available, tests may be compared based on their expected utility (see Section 8.4.4).

8.7 Inference

The calculations of the evaluation measures of accuracy studies as e.g., for sensitivity and specificity present only estimations of the true parameters. As our calculations are only based on the study population which is only a random sample of the target population the values of the parameters vary between different study population (see Figure 8.7).

To allow inference on the true parameter values in the target population, 95%-confidence intervals can be calculated. They describe the precision of the estimate and thus help to assess the uncertainty associated with an estimate. A 95%-confidence interval is always defined such that at least 95 % of all confidence intervals we compute cover the true parameter. As confidence intervals are not uniquely determined, for most parameters there exist a number of methods for calculating such an interval.

8.7.1 General principle

For estimation of a confidence interval, one needs a point estimate of the parameter itself (for example, sensitivity), here denoted $\hat{x}$, and an estimate of its standard error, $\text{SE}(\hat{x})$. The standard error is the square root of the estimated sampling variance of the estimate, $\text{SE}(\hat{x}) = \sqrt{\text{Var}\,(\hat{x})}$. These variances depend on the type of estimator. That is, we need a formula for the variance of

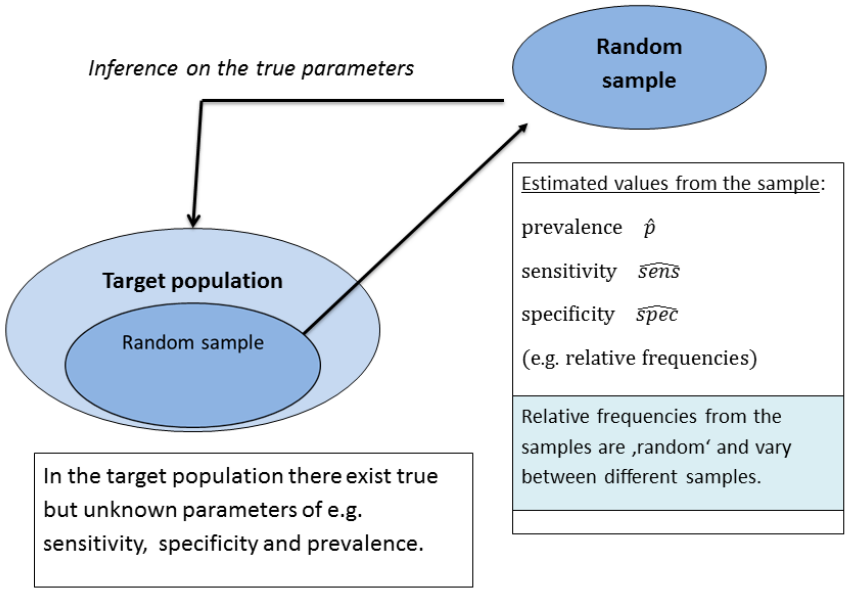


Figure 8.7: Inference

the estimate. Given such a formula and using normal approximation, a 95% confidence interval is calculated by

$$[\hat{x} - 1.96 \times \mathsf{SE}(\hat{x}), \hat{x} + 1.96 \times \mathsf{SE}(\hat{x})]$$

where 1.96 is a constant coming from the normal distribution.

For some measures the parameter is log-transformed, the standard error is computed on the logarithmic scale, and the confidence limits are then back-transformed to the original scale by using the exponential function. This can (but need not) be done for probabilities, such as sensitivity and specificity. It is always done for ratios of probabilities and odds ratios. For these, a 95% confidence interval on the log scale is calculated by

$$[\ln \hat{x} - 1.96 \times \mathsf{SE}(\ln \hat{x}), \ln \hat{x} + 1.96 \times \mathsf{SE}(\ln \hat{x})]$$

and back-transformation to the original scale provides

$$[\hat{x} \times \exp(-1.96 \times \mathsf{SE}(\ln \hat{x})), \hat{x} \times \exp(1.96 \times \mathsf{SE}(\ln \hat{x}))]$$

where $\exp$ denotes the exponential function, $\exp(x) = e^x$.

Sometimes instead of the natural logarithm another transformation is used, the logit (or log odds) transformation. This is defined as follows:

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln p - \ln(1-p)$$

This means that any number p between 0 and 1 is first transformed into its 'odds', which is $p/(1-p)$ and lies between zero and infinity. Then it is log-transformed, and the result can be any real number, including plus and minus infinity. The value can be back-transformed using the inverse of the logit function, often called expit and defined by

$$\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{e^x}{1+e^x}.$$

We will come back to this transformation in Chapter 11. Confidence intervals on the logit scale can be backtransformed to the original scale by applying the expit function to the confidence interval limits. Other common transformations are the arcsine transformation and the Freeman-Tukey arcsine transformation (Anscombe, 1948; Freeman and Tukey, 1950).

In the next subsection we provide formulas for the variance for single probabilities, ratios of independent probabilities and odds ratios. 95% confidence intervals can then be obtained by the general principles as described. We note that there exist further recommended methods for obtaining confidence intervals that are not based on transformations, such as Wilson's score method (Newcombe, 1998).

8.7.2 Variance estimation

Variances for probabilities

For probabilities (such as sensitivity, specificity, prevalence and predictive values) there exists a large number of proposals for estimating the variance (Newcombe, 1998). The simplest is the so-called Wald method: the variance is obtained by taking the estimated values, here denoted $\hat{p}$, and the sample size of the study group, here denoted n, and calculating

$$\text{Var}\,(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}.$$

This simple formula for the variance (Newcombe, 1998, providing also others) comes from the normal approximation to the binomial distribution. As an example, we take the sensitivity:

$$\text{sens} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

The variance is estimated by

$$\text{Var}\,(\text{sens}) = \frac{\text{sens} \times (1 - \text{sens})}{\text{TP} + \text{FN}}$$

which can also be written

$$\text{Var}\,(\text{sens}) = \frac{\text{TP} \times \text{FN}}{(\text{TP} + \text{FN})^3}.$$

As an example, we take from Table 8.8:

$$\text{sens} = \frac{54}{192} = 0.281$$

$$\text{Var}\,(\text{sens}) = \frac{54 \times 138}{192^3} = 0.001.$$

$$\text{spec} = \frac{272}{300} = 0.907$$

$$\text{Var}\,(\text{spec}) = \frac{54 \times 138}{192^3} = 0.00028.$$

A disadvantage of the Wald method is that confidence intervals based on the Wald variance may exceed 1 if the probability is near 1, which clearly does not make sense (likewise, confidence intervals near 0 may contain negative values). To avoid this, the confidence interval may either be accordingly truncated, or the problem can be avoided by appropriately transforming the probability to another scale, determing a confidence interval, and back-transforming to the original probability scale. For transforming, several proposals have been made. We demonstrate the procedure by using the logit transformation which was defined above. For the variance of the logit-transformed probability we use the formula

$$\text{Var}\,(\text{logit}\ \hat{p}) = \frac{1}{n\hat{p}} + \frac{1}{n(1 - \hat{p})}.$$

For the sensitivity and specificity examples from above we obtain

$$\text{logit sens} = \text{logit}\ \frac{\text{TP}}{\text{TP} + \text{FN}} = \ln \text{TP} - \ln \text{FN} = \ln 54 - \ln 138 = -0.9383$$

$$\text{logit spec} = \text{logit}\ \frac{\text{TN}}{\text{TN} + \text{FP}} = \ln \text{TN} - \ln \text{FP} = \ln 272 - \ln 28 = 2.273598$$

and for the variances

$$\text{Var}\,(\text{logit sens}) = \frac{1}{\text{TP}} + \frac{1}{\text{FN}} = \frac{1}{54} + \frac{1}{138} = 0.0258$$

$$\text{Var}\,(\text{logit spec}) = \frac{1}{\text{TN}} + \frac{1}{\text{FP}} = \frac{1}{272} + \frac{1}{28} = 0.03939.$$

Note that these are the variances of the logit-transformed probabilities, therefore they need not be similar, let alone equal to the variances of the probabilities given above. They have no interpretation for their own, but we will make use of them for computing confidence intervals below.

Variances of ratios of independent probabilities

These are estimated using another way to compute the variance using the log transformation. We first give the variance of $\ln p$,

$$\text{Var}\,(\ln \hat{p}) = \frac{1}{n\hat{p}} - \frac{1}{n}.$$

For a ratio of two independent probabilities p and q we use

$$\text{Var}\left(\ln \frac{p}{q}\right) = \text{Var}\,(\ln \hat{p} - \ln \hat{q}) = \text{Var}\,(\ln \hat{p}) + \text{Var}\,(\ln \hat{q}) = \frac{1}{n_1\hat{p}} - \frac{1}{n_1} + \frac{1}{n_2\hat{q}} - \frac{1}{n_2}$$

We may apply this to the positive likelihood ratio (LR+) from Section 8.4.1, again using Table 8.8 as an example. LR+ is estimated by

$$\text{LR+} = \frac{\text{TP}/(\text{TP}+\text{FN})}{\text{FP}/(\text{FP}+\text{TN})} = \frac{\text{TP}(\text{FP}+\text{TN})}{\text{FP}(\text{TP}+\text{FN})} = \frac{54(28+272)}{28(54+138)} = \frac{54\times 300}{28\times 192} = 3.01$$

$$\ln \text{LR+} = \ln\left(\frac{54\times 300}{28\times 192}\right) = 1.103$$

With the sample sizes $n_1 = \text{TP}+\text{FN}$ and $n_2 = \text{FP}+\text{TN}$ the variance formula provides

$$\text{Var}\,(\ln \text{LR+}) = \frac{1}{\text{TP}} - \frac{1}{\text{TP}+\text{FN}} + \frac{1}{\text{FP}} - \frac{1}{\text{FP}+\text{TN}} = \frac{1}{54} - \frac{1}{192} + \frac{1}{28} - \frac{1}{300} = 0.0457.$$

Variance of the diagnostic odds ratio

The variance for the diagnostic odds ratio (DOR) is again estimated via the logit transformation. For the example in Table 8.8 we have

$$\text{DOR} = \frac{54\times 272}{28\times 138} = 3.80$$

$$\ln \text{DOR} = \ln \frac{54\times 272}{28\times 138} = 1.335$$

with estimated variance

$$\text{Var}\,(\ln \text{DOR}) = \text{Var}\,(\text{logit sens}) + \text{Var}\,(\text{logit spec}) = \frac{1}{\text{TP}} + \frac{1}{\text{FP}} + \frac{1}{\text{FN}} + \frac{1}{\text{TN}}$$

which for the example provides

$$\text{Var}\,(\ln \text{DOR}) = \frac{1}{54} + \frac{1}{28} + \frac{1}{138} + \frac{1}{272} = 0.0652.$$

8.7.3 95% confidence intervals for the examples

Sensitivity and specificity (using Wald method)

With sens $= 0.281$ and Var (sens) $= 0.001$ we obtain for the 95% confidence interval of the sensitivity

$$0.281 \pm 1.96 \times \sqrt{0.001} = [0.218; 0.345]$$

and with spec $= 0.907$ and Var (spec) $= 0.00028$ we obtain for the 95% confidence interval of the specificity

$$0.907 \pm 1.96 \times \sqrt{0.00028} = [0.874; 0.940].$$

Sensitivity and specificity (via logit transformation)

With logit sens $= -0.9383$ and Var (logit sens) $= 0.0258$ we obtain for the 95% confidence interval of the logit sensitivity

$$-0.9383 \pm 1.96 \times \sqrt{0.0258} = [-1.2529; -0.6237]$$

Back-transforming this using the expit function, we obtain

$$0.281\ [0.222; 0.349]$$

for the sensitivity and its 95% CI. With logit spec $= 2.2736$ and Var (logit spec) $= 0.0394$ we obtain for the 95% confidence interval of the logit specificity

$$2.2736 \pm 1.96 \times \sqrt{0.0394} = [1.8846; 2.6626].$$

Back-transforming this using the expit function, we obtain

$$0.907\ [0.868; 0.935]$$

for the specificity and its 95% CI.

Positive likelihood ratio

For the positive likelihood ratio, we obtain the 95% confidence interval

$$\exp(\ln 3.01 \pm 1.96 \times \sqrt{0.0457}) = [1.982; 4.582].$$

Diagnostic odds ratio

For the diagnostic odds ratio, we obtain the 95% confidence interval

$$\exp(\ln 3.80 \pm 1.96 \times \sqrt{0.065}) = [2.305; 6.269].$$

Remark For calculating the confidence intervals, we have used that sensitivity and specificity for a fixed cutoff are independent (otherwise it would have become necessary to include a covariance term). This can be justified because the diseased and the disease-free individuals are independent populations. If, however, we would aim at estimating a so-called confidence band for a whole ROC curve, we would need to adjust for the dependence of multiple pairs of sensitivity and specificity due to their negative association when computed from the same population.

Summary of Chapter 8

Measures of diagnostic test accuracy such as sensitivity, specificity and predictive values are based on the theoretical concept of conditional probability. Sensitivity and specificity are the crucial measures of accuracy, while the predictive values also depend on the prevalence of the target condition in the study population. Further measures, derived from sensitivity and specificity, are the likelihood ratios and the diagnostic odds ratio. Moreover, given information about the prevalence of the target condition and costs and/or utilities of true and false test results, the expected utility of the test for diagnosis of the target condition in a specific population can be assessed.

The ROC curve is useful to assess the suitability of a continuous marker as a diagnostic instrument, and the area under the curve (AUC) is used to measure it. To obtain a dichotomous decision rule, a cutoff has to be defined. To this aim, it is important to account for the intended scope of application of the test.

Part D

Special Topics

Introduction Part D

In part A we have considered some basic issues in planning diagnostic accuracy or benefit studies. In part B we have given an overview about actual options for conducting diagnostic accuracy or benefit studies. Part C listed formulas for measures of diagnostic accuracy.

In part D we discuss various remaining issues: Chapter 9 treats sample size considerations for accuracy studies. In Chapter 10, we discuss various aspects of choosing the correct study design. There follows an overview at meta-analysis of diagnostic accuracy studies (Chapter 11). Chapter 12 discusses issues concerning bias, non-inferiority, ethical issues and reporting issues. Finally, in Chapter 13 we venture a look into the future of diagnostic studies.

Chapter 9

Sample Size Considerations

Objectives of Chapter 9

At the end of chapter 9 the reader should be able to ...

- recognize that in accuracy studies actually two sample size calculations are performed.
- recognize that sample size calculations for randomized benefit studies can be performed exactly along the lines applying to all other RCTs

9.1 Sample Size Considerations in Accuracy Studies

Accuracy studies are typically aiming at estimating a pair of parameters: either sensitivity and specificity, or positive and negative predictive value. In the following we will focus on the pair sensitivity and specificity, but all considerations apply in a similar manner when considering positive and negative predictive value as primary outcomes. Since an accuracy study aims at estimating sensitivity and specificity (or a difference in these parameters), and since sensitivity is estimated from the subjects for whom the target condition is present and specificity is estimated from the subjects for whom the target condition is absent, we have actually to perform two sample size calculations: We have to determine the number N_1 of subjects for whom the target condition is present, which is necessary to estimate the sensitivity (or the change in sensitivity) with sufficient precision, and the number N_0 of subjects for whom the target condition is absent, which is necessary to estimate the specificity (or the change in specificity) with sufficient precision. If we have determined these numbers N_0 and N_1, then we have to choose the overall sample size N in a way, such that we can expect at least N_1 subjects for whom the target condition is present, and N_0 for whom the target condition is absent. In a case-control design (cf. Section 5.3), we can sample N_1 cases and N_0 controls, such that $N=N_0+N_1$. In all prospective designs, the sample size N depends on the prevalence (prev) of subjects with the target condition. To reach approximately N_1 subjects with the target condition, we need that prev$\times$N is equal to N_1, or, in other words, the overall sample size N must be at least N_1/prev. To reach approximately N_0 subjects for whom the target condition is absent, we need that $(1-\text{prev})\times N$ is at least equal to N_0, or, in other words, the overall sample size N must be at least $N_0/(1-\text{prev})$. So the overall sample size N is the maximum of N_1/prev and $N_0/(1-\text{prev})$. As the prevalence is typically not exactly known prior to a study, we can often only specify an interval $[\text{prev}_{\min},\text{prev}_{\max}]$, in which we expect the prevalence to be. If we want to ensure to reach N_0 and N_1 for any possible (empirical) prevalence within this interval, we have to choose N as $\max(N_1/\text{prev}_{\min}, N_0/(1-\text{prev}_{\max}))$.

Note that even if $N_0 \sim N_1$, i.e., if we need roughly the same number of subjects with and without the target condition, we may have to enrol much more patients of one of these groups, if the prevalence is not close to 0.5. For example, if we expect a prevalence of 0.2, we need to sample $5\times N_1$ patients to reach the required number N_1. For $4\times N_1$ of these patients the target condition is absent, so here we will sample 4 times the number of patients without the target condition we actually need.

The approach to compute N_0 and N_1 depends on the study design and the method of statistical analysis we want to apply. In single arms studies, results are typically presented as estimates of sensitivity and specificity together with 95% confidence intervals. A formal

hypothesis test is often avoided. In the sample size calculation, we can nevertheless aim at ensuring that the lower limits of the 95% confidence interval are above certain thresholds with a certain probability. For example we can assume a true sensitivity of 90% and try to determine a sample size such that with a probability of 90% the lower bound of the 95% confidence interval will be above 80%. Approximately, this corresponds to a sample size which gives a 90% power to reject the null hypothesis H_0: sens $= 0.8$ if the true sensitivity is 0.9. Hence sample size programs for this type of problems can be used. Table 9.1 shows some sample sizes you obtain by this type of considerations. You can see that if we want to show that the sensitivity or specificity is close to the value we assume, we may need rather large sample sizes.

assumed sensitivity	threshold	sample size N_1
0.70	0.50	62
0.70	0.60	240
0.80	0.60	55
0.80	0.70	200
0.90	0.70	42
0.90	0.80	137
0.95	0.85	96
0.95	0.90	301

Table 9.1: The necessary sample sizes to ensure that the lower bound of the 95% confidence interval is above the given threshold, if the assumed sensitivity is true.

In comparative studies, the statistical approach should be to compute confidence intervals for the difference in sensitivity and in specificity. Significance of such a difference is often tested, too, for example with a McNemar test in paired studies. Sample size calculations can be performed if expectations about the sensitivity and specificity of each test are specified. For example we can specify an expected sensitivity of 80% for the first test and of 90% for the second test, and an expected specificity of 75% for the first test and 85% for the second test, and then we can determine sample sizes N_0 and N_1 such that we have a specific power to find a significant difference. In randomized comparative accuracy studies as consider in Section 5.4 comparison of sensitivity or specificity between the two tests corresponds to a comparison of two proportions between two disjoint groups of subjects. Hence sample size formulas or programs for this situation can be used.

In paired comparative accuracy studies, as considered in section 5.3, the situation is more complicated, as the statistical properties of the estimates depends on the degree of agreement

between the two tests. This degree of agreement is typically not known, and it is not a simple function of the difference in sensitivity and specificity. For example if a test improves the sensitivity from 80% to 90%, the first test and the second test may disagree in 10% of all patients, as only FN→TP changes occur, but it can also happen that they disagree in 30% of all patients, if we have 10% TP→FN changes and 20% FN→TP changes. Since the degree of agreement is not known, we have to start the sample size considerations with some guess, and then we may need to update it later, when we can obtain a first estimate for the degree of agreement. Further considerations about this approach can be found in Alonzo et al. (2002) and Gerke et al. (2008). However, it is an important property of paired accuracy studies that they typically require a distinctly smaller sample size than randomized comparative accuracy studies. This is a simple consequence of the paired nature, such that we can compare test results within each patient, and not only between groups.

In determining N_0 and N_1, it is often useful to use a higher power than the usual 90%. This arises from the fact that we typically want to show that both sensitivity and specificity are improved or above certain levels. Consequently, we should aim at a probability of 90% that both comparisons result in significant results. This implies that the probability to get a significant result for one comparison should be 94.8%, i.e., the square root of 90%.

A further complication in comparative accuracy study may arise from the fact that we expect only an improvement in one parameter, but for the other parameter we only have the hope that it does not become worse. For example we may expect that a new test is able to detect more patients with the target condition, but we do not expect that the test is better to classify patients for whom the target condition is absent. Then we can only expect an improvement in sensitivity, but not in specificity. Consequently, it makes no sense to compare the confidence interval for the change in specificity with 0. We then may decide to analyse specificity like in a non-inferiority study, i.e., we specify a negative number, describing the loss in sensitivity we may be willing to accept, and then compare the lower bound of the confidence interval with this number. The sample size calculation has to be adapted accordingly.

Remark: In Section 7.3 we discussed that it might be wise to analyse paired accuracy studies be linking accuracy to benefit, and that this might be approached by considering weighted averages of the change in sensitivity and specificity. No sample size formulas are available for this approach, but since we overcome the problem of performing separate studies for sensitivity and specificity, we can expect a substantial reduction in sample size. Vach et al. (2012) presented considerations in this direction for single arm studies.

Further reading: A much more detailed account about sample size calculations in accuracy studies can be found in the book by Margaret S Pepe (Pepe, 2003).

9.2 Sample Size Considerations in Benefit Studies

Randomized benefit studies are randomized trials, so sample size calculations can be performed exactly along the lines applying to all other RCTs. The crucial point is to come to a realistic guess for the expected benefit Δ introduced in Section 7.1, as this determines the power and the sample size. The formulas 7.1 and 7.2 of Section 7.1 can help us to get an idea about the benefit we can expect, in particular if we know already from an accuracy study the change in sensitivity and the change in specificity we can expect.

The crucial point here is to understand, why substantial changes in sensitivity and specificity often imply a rather small benefit. Let us start with a simple example: We assume that we can increase both sensitivity and specificity by 20 percentage points. We further assume that both the change from FN to TP as well as the change from FP to TN implies a change in survival probability by 40 percentage points, as we can offer an adequate instead of an inadequate therapy. If we further rely on the reversibility assumption, then formula 7.2 tells us that independent of the prevalence we can expect an overall benefit of $0.2 \times 40 = 8$ percentage points. This is also intuitively clear: An increase in sensitivity and specificity by 20 percentage points implies that we change the test result and hence the treatment decision in 20% of all patients (it may be actually more, but then the changes from for example FP to TN and from TN to FP partially cancel out), and only in these 20% of the patients, i.e., a fifth, we can expect a change in outcome, as we change the treatment. So the 40% advantage in the patients who change their treatment reduces to one fifth of 40%, i.e., 8%.

This reduction can become even more dramatic, if sensitivity and specificity change to a different degree, or if the benefit of FN$\rightarrow$TP changes differs from the benefit of FP$\rightarrow$TN changes, and we have a unfavorable combination of these factors and the prevalence. For example if we have a distinct improvement in sensitivity, but no improvement in specificity, and the prevalence of the target condition is small, only few patient can benefit from the increased sensitivity. Or if mainly patients moving from FN to TP benefit from a better therapy, but the new test mainly increase specificity, we can again only expect a small benefit. Our example in Section 7.3 illustrates also this point: We have a distinct increase in sensitivity, but patients do not benefit a lot when the target condition (distant metastases) is detected, and the prevalence is small, and hence there is at the end no benefit.

Table 9.2 illustrates this point further by some numbers. There we consider two scenarios, namely that changes from FP to TN and from FN to TP benefit to the same degree, or that only changes from FP to TN benefit from the change. We then can observe the relation between the overall benefit and the subgroup specific benefits in dependence on the change in sensitivity, the change in specificity, and the prevalence. We can see how a small prevalence or only a

benefit for FP to TN changes diminishes the benefit if we have mainly a change in sensitivity, and not in specificity. Gated randomized studies considered as Variant 6.1.4 in Section 6.1

c_{sens}	c_{spec}	p	$\Delta/\Delta_{F\to T}$	$\Delta/\Delta_{FP\to TN}$
0.2	0.2	p	0.20	0.10
0.2	0.1	0.2	0.12	0.08
0.2	0.1	0.8	0.18	0.02
0.2	0.0	0.2	0.04	0.00
0.2	0.0	0.8	0.16	0.00

Table 9.2: The expected benefit expressed as a fraction of the subgroup specific benefit in dependence of the change in sensitivity, the change in specificity and the prevalence, based on applying formula 7.2 of Section 7.1. The two scenarios considered are $\Delta_{F\to T} = \Delta_{FN\to TP} = \Delta_{FP\to TN}$ (column 4) and $\Delta_{FN\to TP} = 0$ (column 5).

allow to increase the power and to reduce the necessary sample size, as we remove patients from the analysis, who only add noise. Formulas for sample size calculations for such studies can be found in the paper by Lu and Gatsonis (2013).

Summary of Chapter 9

In accuracy studies actually two sample size calculations are performed, one for subjects for whom the target condition is present and another one for subjects for whom the target condition is absent. For case-control designs, the overall sample size is the sum of the two sample sizes. In prospective accuracy studies, the overall sample size depends on the prevalence of subjects with the target condition. Additional to the study design, determination of the sample size depends on the applied method of statistical analysis. For randomized benefit studies, sample size calculations can be performed exactly along the lines applying to all other RCTs.

Chapter 10

Choosing an Appropriate Design

Objectives of Chapter 10

At the end of chapter 10 the reader should be able to ...

- understand that the choice of the study design is mainly determined by the research question
- recognize that sometimes more than one study design can be appropriate

Although we have considered in Chapters 5 and 6 a lot of design options for diagnostic studies, there are actual few situations where we have really a choice. If you carefully inspect the research questions mentioned for each design, you will realize that the choice of the design is mainly determined be the research question, which reflects the step where we are in the process of developing a new diagnostic test. Accuracy studies are adequate if we want to know the accuracy or if we want to prove that a new test has the potential to improve patient outcomes by improving accuracy. Randomized diagnostic studies are adequate, if we are in doubt about whether we have at the end a benefit for patients by introducing a new diagnostic test in clinical routine. The variants discussed in Section 6.1 are mainly due to specific circumstances. Interaction designs and preselection designs are adequate if we identify a new piece of information by a new test, and if we are in doubt about how this piece of information should influence the treatment decision.

In the following we would like to comment on a few situations where you may have really a choice.

Case-control accuracy studies vs. prospective accuracy studies

The main drawback of case-control studies is the lack of a clear relation to the target population of interest. As cases and controls are drawn from different sources, it remains often unclear whether we can expect similar accuracy estimates if we would perform a prospective study in the target population of interest. Computation of meaningful predictive values from case-control studies is nearly impossible.

The main advantage of case-control studies is the possibility to avoid a long recruitment period, as the patients do not have to be sampled prospectively, but can be taken from an existing patient pool.

Balancing these arguments against each other, it is rather obvious that prospective accuracy studies should be preferred, as they answer the question of interest. case-control studies should be restricted to the early phase of test development, where there may be a need for a small 'proof of principle' study, demonstrating that a test is at least working in the case of very distinct differences between patients with and without the target condition.

Of course, if the prevalence of the target condition is rather low, for example less than 5%, prospective studies may become too expensive to obtain reliable estimates of sensitivity. Then a case-control study might be adequate, if there is a strategy to select cases and controls from one population and avoiding to wait too long until we know the true state of the patients. For example one may follow a large cohort from the target population in which the standard test is applied routinely, apply the new test in a random subsample of for example one percent of the

patients at the same time of the index test, and apply additionally the new test in all cases as soon as they are identified.

Paired comparative accuracy studies vs. randomized comparative accuracy studies

Paired compared accuracy studies offer a lot of advantages compared to randomized comparative accuracy studies, resulting from the simple fact that we apply both tests in each patient. This increases the power substantially and reduces the necessary sample size. Moreover, paired comparative accuracy studies allow us to observe directly the actual changes in test result status which can be helpful in relating accuracy to benefit. They also allow us to restrict the application of the reference status to the patients with discordant results, which may be of interest if the reference standard is expensive or if its application in all patients is ethically difficult to justify (cf. Section 12.3).

A further practical advantage is that paired accuracy studies allow us to study also the accuracy of combinations of the two tests to be compared. For example, we can study the predictive value of the new test, if it is only applied in patients with a positive test result in the standard test, i.e.,, if the new test would be applied as an add-on. Or we can study the number of standard tests we have to apply, if it is only applied in case of a positive result of the new test, i.e., if the new test would be used as a triage test. Such analyses can be very useful, if the overall results for the new test do not suggest using it as a replacement of the standard test.

Due to the many advantages of paired accuracy studies compared to randomized accuracy studies, the use of the latter should be restricted to the case where it is impossible for practical or ethical reasons to apply both tests in each patient.

Gated randomized diagnostic studies

The basic idea of gated randomized studies is to increase the power by analysing only the patients with discordant test results, as all other patients actually add only noise to the estimate of the benefit. The gain can be substantial, so when designing a randomised diagnostic study, this option should be taken into consideration.

However, gated studies require applying the two tests to be compared in each patient, i.e., a sometimes substantial higher burden for the patients and the budget of the study. Moreover, additionally difficulties arise from the question, how and to which degree we inform the patient and the treating physician about the test result. Should we inform them only about that one of the two tests was positive, but not telling them which one? In any case, in a gated study we cannot mimic exactly the clinical routine situation later, when only one test will be performed. So this requires considering the impact on the generalizability when planning a gated study.

Preselection vs. interaction designs

When we are in doubt about whether a positive test result of a new test really suggests offering the patients an alternative treatment, we can choose a preselection design to address this question by comparing the standard treatment with the alternative treatment in patients with a positive test result. However, even if we are successful in demonstrating the superiority of the alternative treatment in such a study, it remains the question whether the alternative treatment may be also beneficial for the patients with a negative test result. Choosing an interaction design allows us exactly to address this question, as we randomize also the test negative patients, too. But it is justified to randomize also these patients, if we actually do not expect any treatment effect in these patients?

This question is hard to answer. It depends on many actual circumstances. Is it really biologically and clinically impossible, that the alternative treatment works in test negative patients, too? If the diagnostic test is based on a biomarker, the test is typically based on some cut-off, and if this cut-off is chosen too high, there may be test negative patients who benefit from the alternative therapy. What will be consequences of not knowing that the alternative treatment is not working in patients with a negative test result? Is there a danger that patients with a negative test result will in future request the alternative treatment, because they may hope that it also beneficial for them? Are there other alternatives for the test negative patients, which are more worth to be tested? (This would suggest also to perform in the test negative patients a separate study.)

Summary of Chapter 10

The design of the diagnostic study is to a great extent determined by the research question. However, in some cases there actually exists a choice between different design options which have to be considered carefully.

Chapter 11

Meta-Analysis of Diagnostic Test Accuracy Studies

Objectives of Chapter 11

At the end of chapter 11 the reader should be able to ...

- explain what are the aims of a meta-analysis of diagnostic test accuracy studies
- interpret the elements of a scatter plot showing the results of a meta-analysis of diagnostic test accuracy studies

Meta-analysis is a statistical method to formally summarize (pool) results from several published or unpublished studies (Borenstein et al., 2009; Higgins and Green, 2011). While the methodology of meta-analysis of intervention studies had been developed and established since the 1980's, modern methods of meta-analysis for diagnostic test accuracy (DTA) studies have been proposed mainly after 2000 (Willis and Quigley, 2011).

11.1 Introduction to Meta-Analysis of DTA Studies

Meta-analysis of DTA studies differs from meta-analysis of intervention studies in a number of aspects. In the following we explain the issues raised by meta-analysis of DTA studies and how to address them (Macaskill et al., 2016).

In a DTA study, the accuracy of a diagnostic test, called the index test, is measured in comparison to a reference test or gold standard (see Chapter 2). Simplified, the data of each DTA study consists of four numbers, the number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). As explained in Chapter 8, sensitivity and specificity of the index test can be estimated from these numbers. We say that the outcome of a DTA study is bivariate, as it consists of the pair of sensitivity and specificity. Some authors prefer to model the pair of true positive rate (which is the sensitivity) and the false positive rate (which is 1 - specificity). Both approaches are completely equivalent.

In a meta-analysis of DTA studies, the pairs of sensitivity and specificity are pooled across a number of DTA studies. However, though the groups of diseased and disease-free individuals are independent within each study, this is not done separately for sensitivity and specificity, as we expect a large across-study correlation between these parameters. This across-study correlation is present because many diagnostic tests are based on an underlying biomarker. If an individuals's value exceeds a certain threshold, the test becomes positive and vice versa. However, we cannot expect all studies to use the same threshold. Studies with a larger threshold probably tend to show a greater specificity and a lesser sensitivity than studies with a smaller threshold. For this reason we often observe a negative correlation between sensitivity and specificity across studies. Therefore, it is not appropriate to conduct separate meta-analyses for sensitivity and specificity (Hamza et al., 2007). Instead, bivariate modelling is necessary.

11.2 Example: Asthma Data

The measurement of fractional exhaled nitric oxide (FeNO) concentration in exhaled air might substitute bronchial provocation for diagnosing asthma. In a systematic review, the diagnos-

tic accuracy of FeNO measurement was investigated compared with the established reference standard (Karrasch et al., 2016). 26 DTA studies with 4578 participants were included. The data are given in Table 11.1 (Karrasch et al., 2016, Table 1).

Study	TP	FN	FP	TN
Arora 2006	87	51	14	20
Cordeiro 2011	33	9	6	66
ElHalawani 2003	7	0	29	13
Florentin 2014	13	6	70	89
Fortuna 2007	17	5	10	18
Fukuhara 2011	33	9	2	17
Giovannini 2014	3	18	0	21
Heffler 2006	14	4	12	18
Katsoulis 2013	23	25	10	54
Kostikas 2008	33	30	13	73
Kowal 2009	157	21	63	299
Linkosalo 2012	13	5	2	10
Malinovschi 2012, current smokers	18	14	14	66
Malinovschi 2012, ex-smokers	12	7	6	37
Malinovschi 2012, never-smokers	35	10	23	40
Pedrosa 2010	26	9	22	57
Pizzimenti 2009	11	3	17	125
Sato 2008	38	10	2	21
Schleich 2012	29	53	4	88
Schneider 2013	75	79	60	179
Sivan 2009	59	10	5	39
Smith 2004	15	2	6	24
Smith 2005	15	12	2	23
Tilemann 2011	25	61	10	114
Voutilainen 2013	13	17	6	51
Wang 2015, bronchodilatation	134	51	83	247
Wang 2015, bronchoprovocation	65	60	16	327
Woo 2012	95	72	10	68
Zhang 2011	29	10	9	58

Table 11.1: Data of the asthma review. TP denotes true positives, FN false negatives, FP false positives and TN true negatives.

11.3 Methods for Meta-Analysis of DTA Studies

11.3.1 Scatterplot of sensitivity and specificity

We assume that each study reports only the numbers TP, FN, TN and FP, as in Table 11.1. That means that there is only one pair of sensitivity and specificity per study. As said before, statistical models for DTA studies model the bivariate distribution of sensitivity and specificity. Before fitting such a model, we look at a scatterplot of sensitivities and specificities. Such a plot is shown for the data example in Figure 11.1, based on the sensitivities and specificities given in Table 11.2. The gray lines correspond to confidence regions of the primary studies. We observe large heterogeneity of sensitivities and false positive rates across studies. We also see the large uncertainty in the results, represented by the confidence regions, due to many studies being small.

Heterogeneity between DTA studies is to be expected, and has two principal causes (Rutter and Gatsonis, 2001). The first source of heterogeneity is variation between DTA studies in the threshold value used to dichotomize the underlying measure into a test result. If there was no other source of heterogeneity, this would lead to a number of points from a single ROC curve common to all the studies. Apparently, this is not the case in our example. However, at least we see a positive correlation between sensitivity (which is the true positive rate) and the false positive rate, but with additional large variation. Hence we conclude that there is a second source of heterogeneity: accuracy probably varies between studies due to clinical heterogeneity in patient populations and/or differences in the implementation of the diagnostic test. For this reason, the observed points need not lie on a common ROC curve.

11.3.2 Models for meta-analysis of diagnostic test accuracy studies

The information shown in Table 11.1 and illustrated in Figure 11.1 provides the basis for statistical analysis of DTA studies. Such meta-analyses may have several aims. First, we want to estimate an average sensitivity and specificity with a joint confidence region. In addition, we may wish to estimate a prediction region where future pairs are expected to be found. Finally, we are interested in a summary ROC (SROC) curve across the observed studies.

For these goals statistical modelling is necessary. Two models have become established during the last decade, a hierarchical model (Rutter and Gatsonis, 2001) and a bivariate model (Reitsma et al., 2005). However, as two groups of researchers independently showed (Harbord et al., 2007; Arends et al., 2008), the hierarchical and bivariate models are equivalent in the special (and most common) case of absence of covariates. In this chapter, we restrict ourselves

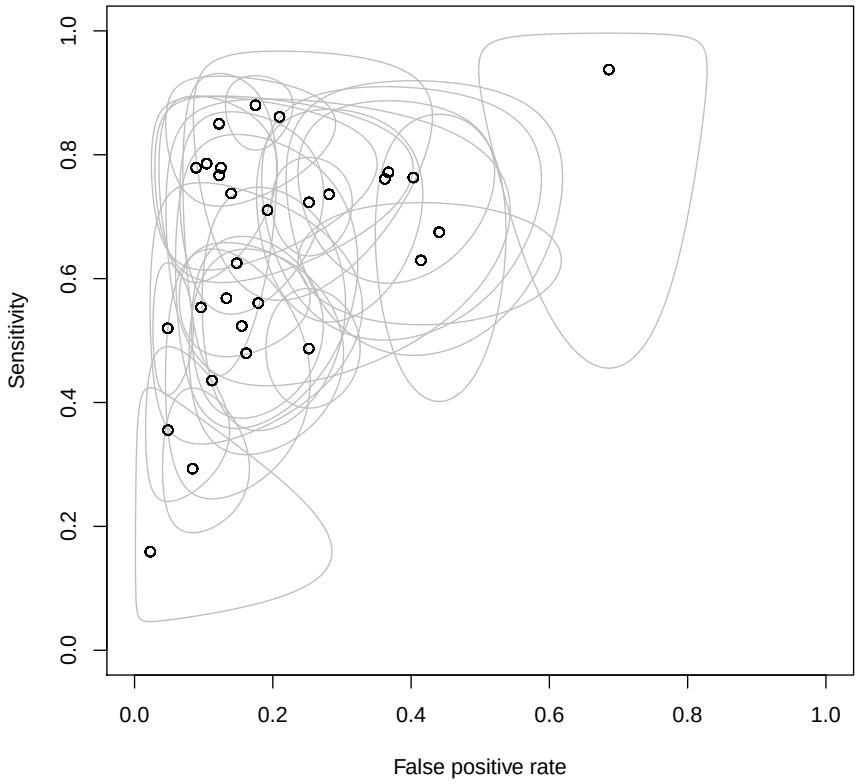


Figure 11.1: Scatter plot of (1 − specificity) and sensitivity with confidence regions for the asthma data.

to the bivariate model.

The bivariate model Reitsma et al. (2005) proposed a bivariate model of the joint distribution of sensitivity and specificity, allowing for across-study correlation. This two-level model followed an approach earlier developed for meta-analysis of binary outcomes (van Houwelingen et al., 1993). It was then refined and generalized by others (Chu and Cole, 2006; Arends et al., 2008).

Primary study level At the first level, for each primary study k, $k = 1, \ldots, K$, we make an assumption how the true positives and false positives within the study are distributed. The underlying idea is that, for each individual in the group, the diagnostic test is thought as a random experiment, where the correct test result appears with probability sens_k (for the

Study	sensitivity	specificity
Arora 2006	0.629	0.586
Cordeiro 2011	0.779	0.911
ElHalawani 2003	0.938	0.314
Florentin 2014	0.675	0.559
Fortuna 2007	0.761	0.638
Fukuhara 2011	0.779	0.875
Giovannini 2014	0.159	0.977
Heffler 2006	0.763	0.597
Katsoulis 2013	0.480	0.838
Kostikas 2008	0.523	0.845
Kowal 2009	0.880	0.825
Linkosalo 2012	0.711	0.808
Malinovschi 2012, current smokers	0.561	0.821
Malinovschi 2012, ex-smokers	0.625	0.852
Malinovschi 2012, never-smokers	0.772	0.633
Pedrosa 2010	0.736	0.719
Pizzimenti 2009	0.767	0.878
Sato 2008	0.786	0.896
Schleich 2012	0.355	0.952
Schneider 2013	0.487	0.748
Sivan 2009	0.850	0.878
Smith 2004	0.861	0.790
Smith 2005	0.554	0.904
Tilemann 2011	0.293	0.916
Voutilainen 2013	0.435	0.888
Wang 2015, bronchodilatation	0.723	0.748
Wang 2015, bronchoprovocation	0.520	0.952
Woo 2012	0.568	0.867
Zhang 2011	0.738	0.860

Table 11.2: Diagnostic accuracies for the asthma data.

diseased) and spec_k (for the disease-free) in study k. Following the proposal by Chu and Cole (2006), we assume that the true unknown sensitivity in study k is sens_k and the true unknown specificity in study k is spec_k. Note that both parameters depend on k, that is, they are specific for the study. The observed numbers TP_k and FP_k each are assumed to follow a

binomial distribution with parameters $(\mathsf{R}_k^+, \mathsf{sens}_k)$ and $(\mathsf{R}_k^-, \mathsf{spec}_k)$, respectively, where R_k^+ and R_k^- are the numbers of true diseased and true disease-free individuals in study k.

Across-study level At the second level, we make an assumption how the true unknown sensitivities and specificities are distributed across the studies in the meta-analysis. The sensitivities and specificities are probabilities, lying in the interval $[0, 1]$. In statistics, probabilities often are transformed such that they can assume any real value. (This is the case, for example, when using the widely known logistic regression model that allows to make predictions on event probabilities.) Also in the present application, sensitivities and specificities are transformed using the so-called logit (or log odds) transformation. This is defined as follows (see also Chapter 8):

$$\mathsf{logit}(p) = \ln \frac{p}{1-p} = \ln p - \ln(1-p)$$

A number p between 0 and 1 is first transformed into its 'odds', which is $p/(1-p)$ and lies between zero and infinity. Then it is log-transformed, and the result can be any real number, including plus and minus infinity.

Here, we use the logit-transformed sensitivities and false positive rates. They correspond to a scatter plot which is seen in Figure 11.2. This scatter plot (where the large across-study heterogeneity is seen again) is modeled using a bivariate normal distribution. It has five parameters that must be estimated:

- two means, μ_1 for logit(1 - specificity) and μ_2 for logit(sensitivity)
- two variances σ_1^2, σ_2^2,
- a correlation coefficient ρ between both, which in our example is (but in general does not need to be) positive.

As a model equation, this looks like this:

$$\begin{pmatrix} \mathsf{logit}(\mathsf{sens}_k) \\ \mathsf{logit}(1-\mathsf{spec}_k) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

There are a number of software packages for fitting such models. In the open statistical environment R (R Core Team, 2014), there are the R packages mada (Doebler, 2015) and meta4diag, which uses a Bayesian approach (Guo and Riebler, 2015). A Stata package to this aim is metandi (StataCorp., 2013). We have analyzed the asthma data using the R package mada and present the result in Section 11.4.

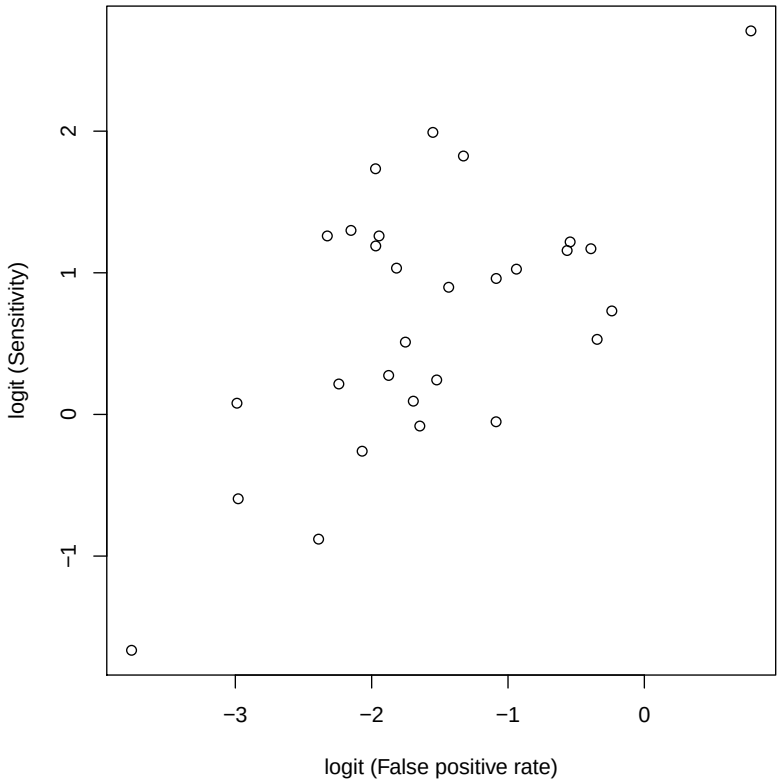


Figure 11.2: Scatter plot of logit(1 − specificity) and logit(sensitivity) with confidence regions for the asthma data.

11.3.3 Methods for estimating a summary ROC curve

We have seen a scatterplot of false positive rates (1 − specificities) versus the true positive rates (sensitivities), see Figure 11.1. Now we are interested in estimating a summary ROC curve across the studies. The simplest idea would be to regress the logit-transformed TPRs against the logit-transformed FPRs and then back-transform to the ROC space. However, this curve addresses only one very specific question, namely, 'How large is the sensitivity, given the specificity?'. As usual in regression, exchanging sensitivity and specificity gives a different curve. Neither curve is symmetric with respect to sensitivity and specificity and both can be misleading (Rücker and Schumacher, 2009).

Moses et al. (1993) proposed a summary ROC curve based on a regression of the difference of the logit-transformed positive rates (that is, the logarithm of the diagnostic odds ratio)

against their sum (which is a proxy for the threshold, see (Arends et al., 2008, eq. (15))). It is symmetric with respect to sensitivity and specificity, but does not account for potential heterogeneity or for the different precision of the estimates from different studies. The proposal by Rutter and Gatsonis (2001) leads to a different solution. The slope of this line in the logit space is the geometric mean of the slopes of the two regression lines, $\text{logit}(\text{sens})$ on $\text{logit}(1 - \text{spec})$ and vice versa (Arends et al., 2008, eq. (16)).

Arends et al. pointed out that the SROC curve is in principle unidentifiable if only one $(\text{sens}, \text{spec})$ pair per study is known (Arends et al., 2008).

11.4 Results for the Asthma Data

Figure 11.3 shows the results for the asthma data. Pooling the studies using the bivariate model provided a pooled sensitivity of 0.65 with a 95% confidence interval 0.58 to 0.72 and a pooled specificity of 0.82 (95% confidence interval 0.76 to 0.86). This point estimate is represented by a black dot.

The bivariate (two-dimensional) 95% confidence region is shown as a solid circle. This is a statement on the pooled estimate. Its interpretation is that if we perform 100 such meta-analyses, each with a new set of the same type of data, the true unknown pair of sensitivity and specificity will be covered by 95 of the resulting confidence regions. The dotted circle gives the 95% prediction region. This is a statement on future studies. It means that we expect that 95 of 100 future studies will probably lie within the prediction region.

The bold solid curve provides a summary ROC curve, estimated following the above-mentioned proposal by Rutter and Gatsonis (2001), see also Macaskill (2004); Harbord et al. (2007).

11.5 Further issues

We have given a short overview of the issues raised in meta-analysis of diagnostic test accuracy studies, which may be seen as a special case of multivariate meta-analysis. We concentrated on the bivariate model, which provides an estimate of the pooled sensitivity and specificity with confidence region and prediction region. Of the many proposals for estimating a summary ROC curve, we chose that proposed by Rutter and Gatsonis.

It must be noted that all methods we have described here are based on the assumption that there is only one pair of sensitivity and specificity known per study. The underlying threshold was ignored and not used in the models. This is a shortcoming of the bivariate and other

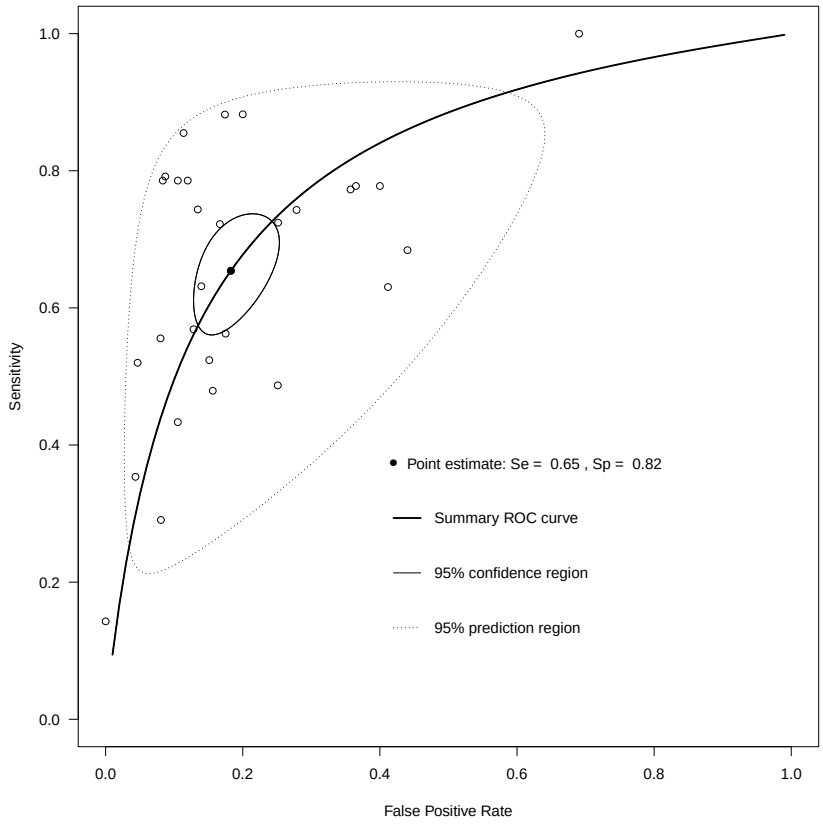


Figure 11.3: ROC plot for the asthma data. Bold black dot: Pooled estimate of sensitivity and specificity. Solid circle: 95% confidence region. Dotted circle: 95% prediction region. Solid curve: Summary ROC curve.

standard models.

In practice, often in a number of primary studies several pairs are provided and there is also information on the thresholds. Sometimes even full ROC curves are provided. If one wants to make use of this additional information, more complex models are necessary. There are a number of proposals in the literature (Dukic and Gatsonis, 2003; Hamza et al., 2009; Putter et al., 2010; Martínez-Camblor, 2014; Riley et al., 2014, 2015b,a; Steinhauser et al., 2016).

There are also ideas how to compare multiple tests in a so-called network meta-analysis of diagnostic accuracy tests (Trikalinos et al., 2014; Menten and Lesaffre, 2015; Hoyer and Kuss, 2016; Dimou et al., 2016; Nyaga et al., 2016a,b). These methods are beyond the scope of this chapter.

Summary of Chapter 11

Meta-analysis of diagnostic test accuracy studies is a special case of multivariate meta-analysis. We presented the bivariate model, which provides an estimate of the pooled sensitivity and specificity with confidence region and prediction region. We also added a summary ROC curve, as proposed by Rutter and Gatsonis. The methods described here are based on the assumption that there is only one pair of sensitivity and specificity known per study.

Chapter 12

Further issues

Objectives of Chapter 12

At the end of chapter 12 the reader should be able to ...

- recognize that in single arm accuracy studies several types of bias can occur
- differentiate between spectrum bias and verification bias
- recognize that an imperfect reference standard can also produce a bias in single arm accuracy studies
- recognize that diagnostic studies typically require an approval by an ethics committee
- identify the STARD statement, STROBE statement and CONSORT statement as reporting guidelines for diagnostic studies

12.1 Types of Bias in Accuracy Studies

In the literature several types of bias have been discussed which can occur in single arm accuracy studies. We discuss some of these bias types in the light of the design of accuracy studies.

Spectrum bias This term refers to differences in accuracy parameters between different studies due to the fact that they include different clinical populations which differ in accuracy, typically due to differences in the composition of subjects easy to diagnose or difficult to diagnose. Careful considerations about the existence of different target populations and the relation between the actual study population and the target population – as discussed in Section 2.3 – can help to reduce the risk of spectrum bias. It has been discussed (e.g., Mulherin and Miller (2002)), whether the term *spectrum effect* may be more appropriate than the term *spectrum bias*, as we refer here to a property of different populations. However, if one would like to express that the results of a specific study may be not relevant, as the study population is too different from the target population or the intended target population is not clear, the term *spectrum bias* may be appropriate.

Verification bias, workup bias or referral bias These terms refer to the situation that the reference test is performed only for a subgroup of those patients, for whom the index test is performed, and only this subgroup is analysed. This can indeed lead to a bias in sensitivity and specificity, even if the decision for whom the reference test should be performed depends only on the results of the index test. In such a situation a fraction a of the subjects with a positive test result would be tested with the reference test, and a fraction b of the subjects with a negative test result would be tested with the reference test. So the number of TP results observable if we test all subjects would reduce to $a\times$TP , and the number of FN results observable if we test all subjects would reduce to $b\times$FN. Consequently, the sensitivity would change from TP/(TP+FN) to ($a\times$TP)/($a\times$TP + $b\times$FN), which implies a decrease in sensitivity (if $a > b$) or an increase (if $b > a$). Similar considerations can be applied to the specificity.

However, in general we cannot assume that the decision which patients should be sent to the reference test depends only on the results of the index test. The term workup bias and referral bias point to the fact that such decisions are often made in the patient management process, where also other information on the patient is used, for example the patient history or the results of other tests. This entails the risk that only the subjects difficult to diagnose are sent to the reference test, whereas for the subjects easy to diagnose the reference test is omitted. We already discussed this in Section 2.3.

Incorporation bias This term refers to the situation that the index test is part of the reference standard (Whiting et al., 2003). For example, Perucchini et al. (1999) investigated a fasting glucose test for diagnosis of gestational diabetes. The index test was a venous plasma blood glucose measurement, performed in the morning after a 12 overnight fasting aand a diet. The reference standard was the 100g oral glucose test including just the same first venous plasma blood glucose measurement, followed by intake of 100g glucose solution and three measurements of venous plasma blood glucose concentration after 1 hour, 2 hours and 3 hours. Thus the index test was part of the reference standard. This leads to a likely overestimation of diagnostic accuracy (incorporation bias).

Imperfect reference standard This is a further source of bias in single arm accuracy studies. Similar to the bias types considered so far, the impact of an imperfect reference standard can be an overestimation or an underestimation of accuracy. To understand this, let us consider the simple situation that an imperfect reference standard only overlooks some patients for whom the target condition is present, but still correctly classifies all patients for whom the target condition is absent. We here only consider the group of patients with the target condition ('diseased'). We assume an imperfect reference test with true sensitivity $\text{sens}_{ref} = P(R+|D+) =: P_{D+}(R+) < 1$ and a true specificity of one (that is, there are no false positives). Further, we assume an index test with true sensitivity $\text{sens}_{ind} = P(I+|D+) =: P_{D+}(I+)$. We assume that the true sensitivities cannot be observed, because there is no true gold standard. The estimated sensitivity of the index test is

$$\text{sens}^*_{ind} = P_{D+}(I+|R+) = \frac{P_{D+}(I+ \cap R+)}{P_{D+}(R+)}$$

We have

$$\begin{aligned} \text{sens}^*_{ind} > \text{sens}_{ind} \quad &\text{if} \quad P_{D+}(I+|R+) > P_{D+}(I+) \\ \text{sens}^*_{ind} < \text{sens}_{ind} \quad &\text{if} \quad P_{D+}(I+|R+) < P_{D+}(I+) \end{aligned}$$

That is, if (plausibly) there is a positive association between both tests, given the true status, then the sensitivity tends to be overestimated. Only if the conditional correlation is negative (which is less probable), the sensitivity may be underestimated. In other words, if the reference test tends in its imperfectness to become closer to the index test, we overestimate accuracy, and if it tends to move away from the index test, we underestimate accuracy. Often it is possible to get an idea about this when considering the structure of the index test and the reference test. If, for example, there is a risk that the reference test still overlooks patients with mild

symptoms or in a very early stage of disease, and we have the same fear for the index test, then we may expect an overestimation of the accuracy due to using an imperfect reference standard.

It should be emphasized that all these considerations about bias refer to single arm studies and the impact on accuracy parameters. These considerations do not automatically apply to comparative accuracy studies, as there the differences in accuracy parameters are the quantities of interest. Roughly speaking, if both tests to be compared in a comparative study suffer to the same degree from some bias, the difference may be still rather unbiased. Unfortunately, until now there exist no systematic investigations about the impact of partial verification or imperfect reference standards on the difference in accuracy parameters, so it is too early to give here an overview about the potential impact. However, it can be already said that one crucial aspect with respect to the impact of an imperfect reference standard will be the question, whether an imperfect reference standard may be more similar to one of the two tests to be compared than to the other. For example, if the reference standard is based on a follow up, and in the follow up one of the two tests to be compared is used again, it should not be surprising that this test also yields better accuracy.

12.2 Diagnostic Tests with a Direct Benefit for the Patient

So far we have assumed that the benefit for a patient from a new diagnostic test arises only due to an improved accuracy. There are situations where a new diagnostic test offers a direct benefit for the patient. Replacement of an invasive test by a non-invasive test is a typical example of this kind, for example if we can replace a surgery with investigating the status of the patient by an imaging technique. Other situations arise if a new test is less time consuming or if it can be done without laboratory equipment, such that it can be applied also in settings where no laboratory is available. A test may be also of interest from a societal perspective, if it is just much cheaper.

In these situations it is sufficients for an accuracy study to demonstrate that the benefit from the test is not counterbalanced by a decrease in accuracy. So we need a comparative accuracy study comparing the new test with the current standard. However, this study should not aim at showing an improvement in accuracy, but just that the accuracy is comparable. Typically, we will require that the differences in sensitivity, in specificity and in the predictive values are small, even if we take the boundaries of the confidence intervals into account.

However, it may be argued that this is not sufficient, as similar sensitivities and specificities may appear even if the results of the two tests differ in many patients, and that differences

occur in both directions and hence imply finally similar sensitivities and similar specificities. This has to be taken seriously, as it may imply that the patient populations with positive test results or negative test results change, and we cannot guarantee that these populations will still benefit from the established management processes. So it would be clever to perform the comparison by a paired study and to report not only the change in the accuracy parameters, but also the actual individual agreement of test results between the two tests.

12.3 Ethical issues in diagnostic studies

As diagnostic studies involve research in humans, they typically require an approval by an ethics committee. As the rules applied by ethics committees vary from country to country and often also within countries, the requirements made by ethics committees for diagnostic studies are not necessarily uniform.

In an accuracy study, a typical question to be addressed to obtain an approval by the ethics committee is whether the index test(s) and reference test to be used are already part of the clinical routine, or if they are applied in addition to the current practice. In the latter case, the ethics committee may raise some concern, in particular, if the additional tests to be applied are invasive or imply an exposure to radiation. Then it may be important to emphasize that even if the tests have to be performed in a blinded manner, the results of the tests can be communicated to the patient and the treating physician as soon as all test results of the patient have been determined. So patients can often have a direct benefit from the additional tests, even if the results have to be kept secret for a (short) time period.

A crucial point for an ethics committee can be the wish of the investigator to apply a reference test also for subjects with negative test results for both tests in a comparative accuracy study. Often these subjects do not require any further action from a clinical point of view, and it would be hard to justify additional investigations like a biopsy. Here it might be wise to think of whether this is really necessary. We have pointed out in Section 7.3 that paired accuracy studies can be analysed in a way only requiring the reference test for subjects with discordant results, so this way we can avoid this problem. It is also possible to estimate the ratio of the sensitivity between the two tests and the ratio of the specificity between the two tests without knowing the reference test result in subjects with negative test results in both tests, as pointed out by Schatzkin et al. (1987).

What remains in any case is the need to know the reference test results for patients with discordant results. Even this may be problematic for an ethics committee, if in clinical routine the reference test (e.g., surgery) is only applied in patients with a positive test result in the

standard test. Here it will be necessary to argue that the number of patients with a negative test result in the standard tests and a positive test result in the new test, i.e., those for whom we require additionally to apply the reference test, is typically rather small, and that these additional tests are necessary to make any statement about the accuracy of the new test. In the long run, this is in the interest of all patients, as it may be a first step in finally justifying to switch to the new test.

Until now we do not have enough experience with randomized diagnostic studies to identify potential issues in getting an approval from an ethics committee. Typically, they will require that we are in the state of equipoise, i.e., that we do not have sufficient evidence for an advantage of one of the two diagnostic procedures to be compared. This may require to perform a systematic literature research about comparative diagnostic studies comparing the two procedures, covering both accuracy and benefit studies (and perhaps also other study types) and a meta-analysis of their results, see Chapter 11.

12.4 Reporting

The best planning and conduct of a diagnostic study is worthless, if the results are not published such that all essential properties, strengths (and limitations) are communicated in an adequate manner, to enable others to judge the generalizability and value of the results. Careful selection of the study population, sophisticated construction of a reference test, blinding of all tests, standardization of the tests and of the management procedures etc. can substantially contribute to the value of a study, but these points have to be reported in the publication. Unfortunately, it is often forgotten to report all these essential details. Fortunately, today there exist reporting guidelines, which can assist in an adequate reporting of diagnostic studies. The most relevant are the STARD statement (STAndards for the Reporting of Diagnostic accuracy studies; Bossuyt et al. (2003)) with respect to accuracy studies, but for these studies it might be also worth to take a look at the STROBE statement for observational studies in general (von Elm et al., 2007). For randomized benefit studies, it is worth to consider the CONSORT statement for the reporting of clinical trials (Schulz et al., 2010) and its variants for non-phamacological treatment interventions (Boutron et al., 2008) and pragmatic trials (Zwarenstein et al., 2008).

However, lack of reporting of important design issues in diagnostic studies may also be due to lack of space in the main publication. Hence it might be worthwile to consider publishing the study protocol separately in order to allow a complete description of the design. Many open access journals today support the publication of study protocols, and they may even accept them without starting a review process, if the study itself was subjected to an external review

when applying for funding.

Summary of Chapter 12

In single arm accuracy studies several types of bias can occur. Spectrum bias refers to differences in accuracy parameters between different studies due to different clinical populations which differ in accuracy. Verification bias (workup bias or referral bias) refers to the situation that the reference test is performed only for a subgroup of those patients, for whom the index test is performed, and only this subgroup is analysed. An imperfect reference test does not present the truth and can cause an overestimation or underestimation of the accuracy.

Chapter 13

The Future of Accuracy and Benefit Studies

Objectives of Chapter 13

At the end of chapter 13 the reader should be able to ...

- explain the differences between diagnostic accuracy trials, diagnostic benefit studies and pure intervention trials
- recognize that only a small number of participants in a diagnostic benefit study may actually benefit from better diagnosis
- understand why diagnostic benefit studies need large sample size

As mentioned in the preface, we experience today a great transition in the field of diagnostic research. Guideline developers and health policy makers require evidence for a patient benefit, and this has fundamentally questioned the dominating role of accuracy studies. Randomized studies have been advocated as the tool of the future. So at the end of this course, we may try to take a look into the future.

Accuracy studies will definitely still play an important role in the future, as long as we continue developing new tests trying to improve patient benefit by more accurate diagnoses. For such tests successful accuracy studies will be always a prerequisite for any randomized diagnostic study, as without any evidence for accuracy, it would be hard to justify a randomized study. Accuracy studies providing a comparison with the current standard test should be regarded as the standard study type, as they provide the most valuable information to support any decision about continuing to investigate a new diagnostic test. Single arm studies should be restricted to the case of new diagnostic tests providing a piece of information, where no comparator exists.

It is more difficult to predict the future of randomized benefit studies, in particular whether they will play a role similar to the role of RCTs in therapeutic research. Some HTA bodies like the German IQWiG (Scheibler et al., 2010) propagate the use of a randomized study as the main body of evidence to assess patient benefit. However, randomized diagnostic studies are rather rare to date (Ferrante di Ruffano et al., 2012a). In other countries, HTA bodies are willing to accept the results of other types of studies. For example the Centers for Medicare & Medicaid Services (CMS) in the US seem to be willing to accept already evidence for changes at the level of the patient management as an appropriate basis for decisions (Hillman et al., 2013).

There are some issues which may prevent randomized diagnostic studies from playing the same role as RCTs in therapeutic research.

Sample size Better diagnostic tests only help those patients, who suffer from an incorrect diagnosis based on the current diagnostic standard. So this number is often rather small compared to the whole patient population we apply the test to. This implies that though the individual benefit for a patient with a change in diagnosis can be huge, the overall benefit in the whole patient population may be small. This is a simple consequence of moving the perspective from 'helping those who suffer from an incorrect diagnosis' to 'improving patient benefit on average'. Studies like the MINDACT study (Cardoso et al., 2016) or the ERSPC trial (Schröder et al., 2009) with several thousands of patients may hence be more the rule than the exception, if we want to perform randomized diagnostic studies with sufficient power with patient relevant outcomes.

Choice of outcomes In Section 3.1 we have discussed the different consequences of FP, FN, TP and TN test results, and in Section 8.4.4 we have explained the concept of utilities and costs. The consequences of test results can be of different quality, and it is questionable whether it is reasonable to summarize them always up into one outcome. This may be possible if we distinguish two disease states with curative treatment options, when outcomes like cure rates and survival are available. This is not always the case. In primary diagnosis, we can expect an impact on cure rates and survival only among those who are diseased, but we have to balance this against a loss in quality of life due to false positive test results in the disease-free individuals. If the disease states to be distinguished are associated with the decision between curative and palliative treatment, we have a similar problem.

Standardization The success of RCTs in therapeutic research is mainly confined to the area of developing pharmacological treatments. One reason for this is the ease to standardize most pharmacological treatments: We can use a specific dose of a specific substance to be administered in one specific way. In other fields of therapeutic research, for example surgery or dentistry, RCTs are much less popular. We have pointed out in Section 3 that randomized diagnostic studies are actually evaluating a complex intervention with at least two ingredients, namely a diagnostic procedure and at least two different management processes, and we have also pointed out problems in reaching a sufficient degree of standardization. We have to wait if guideline developers and health policy makers will be willing to accept the results of randomized diagnostic studies, or whether they will question them due to problems with the generalizability, arising from insufficient or artificial standardization.

Test evolution In contrast to most pharmacological treatments, diagnostic tests can evolve outside of trials. The visual resolution of imaging procedures is typically improved every year, and gene expression profiles are improved by adding or combing them with new detected markers. There is a risk that modern diagnostic tests are already outdated, when a large scale randomized study is finished. To prevent this conflict, instead of using a randomized diagnostic study to prove the benefit of a specific test, one could consider it as a means to show the benefit of a more general class of diagnostic tools, independently of the specific test used. For example, we may regard the MINDACT study (Cardoso et al., 2016) less as a tool to show the benefit of the specific 70-gene signature considered, than as a tool to show that using gene profiles for this type of decision can lead to better test results and better patient outcomes compared to conventional criteria. Then it might be possible to use the results of this study also for other gene signatures, which may have shown a higher predictive accuracy for the prognosis of patients outside or randomized diagnostic studies.

On the other hand, we have to be aware that diagnosis and therapy tend to build more and more a single unit in medical research today. Many new biomarkers suggest by the presence of certain molecular features a specific, targeted therapy, and hence there is only one test for one therapy, and the distinction between diagnosing and treatment may vanish. Scheibler et al. (2012) investigated the published and planned randomized studies using PET/CT as a diagnostic modality, and identified in study registers 42 studies. Many of these trials considered the target situations treatment planning and response evaluation, which are two areas where accuracy studies are of limited value. Treatment planning, in particular for radiation therapy, requires to integrate a lot of information into a treatment plan, and hence the results are hard to predict from accuracy studies. Response evaluation with new imaging techniques is often confronted with the problem that patients without a response are offered an alternative treatment, and hence the correctness of the test results cannot be checked. Hence for both types of target situations only the package of the diagnostic test together with the treatment can be evaluated. For those areas where we still develop new diagnostic tests to improve accuracy and not to prepare directly individualized treatment, the lack of randomized studies may remain.

It is hard to predict the future of randomized benefit studies, but it is predictable that in a few years we probably will have a different view on this topic than we have today.

Summary of Chapter 13

Accuracy studies will still play an important role in the future, as long as we continue developing new tests trying to improve patient benefit by more accurate diagnoses. Randomized benefit studies including diagnostic procedures are associated with a number of problems. First, the sample size must be large, as a benefit can only be expected for patients who are incorrectly diagnosed by the standard test. Secondly, the choice of outcomes is an issue, as the consequences for false positives and false negatives are quite different. Thirdly, diagnostic procedures are less standardizable than pharmacological treatments. Finally, diagnostic tests are often developed outside of trials and their evolution can be very fast, meaning that thoroughly planned megatrials can be outdated before their results are available.

Bibliography

Alonzo, T. A., Pepe, M. S., and Moskowitz, C. S. (2002). Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine*, 21(6):835–852.

Alpert, J. S., Thygesen, K., Antman, E., and Bassand, J. P. (2000). Myocardial infarction redefined–a consensus document of the joint european society of cardiology/american college of cardiology committee for the redefinition of myocardial infarction. *Journal of the American College of Cardiology*, 36(3):959–969.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254.

Arends, L. R., Hamza, T. H., van Houwelingen, J., Heijenbrok-Kal, M., Hunink, M., and Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*, 28(5):621–638.

Aviv, J. E. (2000). Prospective, randomized outcome study of endoscopy versus modified barium swallow in patients with dysphagia. *The Laryngoscope*, 110(4):563–574.

Baker, S. G. and Kramer, B. S. (2007). Peirce, Youden, and receiver operating characteristic curves. *The American Statistician*, 61(4):343–346.

Böhning, D., Böhning, W., and Holling, H. (2008). Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical Research*, 17:543–554.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2009). *Introduction to Meta Analysis*. Wiley, Chichester.

Bossuyt, P. and Leeflang, M. (2008). Chapter 6: Developing criteria for including studies. In *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Version 0.4 [updated September 2008].

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., and Standards for Reporting of Diagnostic Accuracy (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. standards for reporting of diagnostic accuracy. *Clinical Chemistry*, 49(1):1–6.

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., Ravaud, P., and CONSORT Group (2008). Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of Internal Medicine*, 148(4):295–309.

Buyse, M., Michiels, S., Sargent, D. J., Grothey, A., Matheson, A., and de Gramont, A. (2011). Integrating biomarkers in clinical trials. *Expert Review of Molecular Diagnostics*, 11(2):171–182.

Cardoso, F., van't Veer, L., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J., Brain, E., Causeret, S., DeLorenzi, M., Glas, A., Golfinopoulos, V., Goulioti, T., Knox, S., Matos, E., Meulemans, B., Neijenhuis, P., Nitz, U., Passalacqua, R., Ravdin, P., Rubio, I., Saghatchian, M., Smilde, T., Sotiriou, C., Stork, L., Straehle, C., Thomas, G., Thompson, A., van der Hoeven, J., Vuylsteke, P., Bernards, R., Tryfonidis, K., Rutgers, E., Piccart, M., and Investigators, M. (2016). 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine*, 375(8):717–729. doi: 10.1056/NEJMoa1602253.

Chu, H. and Cole, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed approach. *Journal of Clinical Epidemiology*, 59:1331–1333.

Clark, N. M., Janz, N. K., Dodge, J. A., Mosca, L., Lin, X., Long, Q., Little, R. J., Wheeler, J. R. C., Keteyian, S., and Liang, J. (2008). The effect of patient choice of intervention on health outcomes. *Contemporary Clinical Trials*, 29(5):679–686.

Cobo, M., Isla, D., Massuti, B., Montes, A., Sanchez, J. M., Provencio, M., Viñolas, N., Paz-Ares, L., Lopez-Vivanco, G., Muñoz, M. A., Felip, E., Alberola, V., Camps, C., Domine, M., Sanchez, J. J., Sanchez-Ronco, M., Danenberg, K., Taron, M., Gandara, D., and Rosell, R. (2007). Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 25(19):2747–2754.

Cutsem, E. V., Köhne, C.-H., Láng, I., Folprecht, G., Nowacki, M. P., Cascinu, S., Shchepotin, I., Maurel, J., Cunningham, D., Tejpar, S., Schlichting, M., Zubel, A., Celik, I., Rougier, P., and Ciardiello, F. (2011). Cetuximab plus irinotecan, fluorouracil, and leucovorin as first-line

treatment for metastatic colorectal cancer: Updated analysis of overall survival according to tumor KRAS and BRAF mutation status. *Journal of Clinical Oncology*, 29(15):2011–2019.

de Graaff, J. C., Ubbink, D. T., Legemate, D. A., Tijssen, J. G. P., and Jacobs, M. J. H. M. (2003). Evaluation of toe pressure and transcutaneous oxygen measurements in management of chronic critical leg ischemia: a diagnostic randomized clinical trial. *Journal of Vascular Surgery*, 38(3):528–534.

De Lorijn, F., Reitsma, J. B., Voskuijl, W. P., Aronson, D. C., Ten Kate, F. J., Smets, A. M. J. B., Taminiau, J. A. J. M., and Benninga, M. A. (2005). Diagnosis of hirschsprung's disease: a prospective, comparative accuracy study of common tests. *The Journal of Pediatrics*, 146(6):787–792.

Dimou, N. L., Adam, M., and Bagos, P. G. (2016). A multivariate method for meta-analysis and comparison of diagnostic tests. *Statistics in Medicine*, 35(20):3509–3523. doi: 10.1002/sim.6919. Epub 2016 Mar 4.

Doebler, P. (2015). mada: Meta-analysis of diagnostic accuracy. http://www.r-project.org/web/packages/mada/. R package version 0.5.7.

Dukic, V. and Gatsonis, C. (2003). Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics*, 59:936–946. doi: 10.1111/j.0006-341X.2003.00108.x.

Eng, K. H. (2014). Randomized reverse marker strategy design for prospective biomarker validation. *Statistics in Medicine*, 33(18):3089–3099.

Ferrante di Ruffano, L., Davenport, C., Eisinga, A., Hyde, C., and Deeks, J. J. (2012a). A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology*, 65(3):282–287.

Ferrante di Ruffano, L., Hyde, C. J., McCaffery, K. J., Bossuyt, P. M. M., and Deeks, J. J. (2012b). Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ*, 344(feb21 1):e686–e686.

Fischer, B., Lassen, U., Mortensen, J., Larsen, S., Loft, A., Bertelsen, A., Ravn, J., Clementsen, P., Høgholm, A., Larsen, K., Rasmussen, T., Keiding, S., Dirksen, A., Gerke, O., Skov, B., Steffensen, I., Hansen, H., Vilmann, P., Jacobsen, G., Backer, V., Maltbæk, N., Pedersen,

J., Madsen, H., Nielsen, H., and Højgaard, L. (2009). Preoperative staging of lung cancer with combined PET–CT. *New England Journal of Medicine*, 361(1):32–39.

Flicker, L., Logiudice, D., Carlin, J. B., and Ames, D. (1997). The predictive value of dementia screening instruments in clinical populations. *International Journal of Geriatric Psychiatry*, 12(2):203–209.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "mini-mental state". a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21:607–611.

Freidlin, B., McShane, L. M., and Korn, E. L. (2010). Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160.

Fryback, D. G. and Thornbury, J. R. (1991). The efficacy of diagnostic imaging. *Medical Decision Making*, 11(2):88–94.

Gazelle, G., Kessler, L., Lee, D. W., McGinn, T., Menzin, J., Neumann, P. J., van Amerongen, D., and White, L. A. (2011). A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. *Radiology*, 261(3):692–698.

Geijerstam, J.-L. a., Oredsson, S., and Britton, M. (2006). Medical outcome after immediate computed tomography or admission for observation in patients with mild head injury: randomised controlled trial. *BMJ : British Medical Journal*, 333(7566):465.

Gerke, O., Hoilund-Carlsen, P., and Vach, W. (2015). Analyzing paired diagnostic studies by estimating the expected benefit. *Biometrical Journal*, 57 (3):395–409.

Gerke, O., Vach, W., and Høilund-Carlsen, P. F. (2008). PET/CT in cancer: Methodological considerations for comparative diagnostic phase II studies with paired binary data. *Methods of Information in Medicine*, 47(6):470–479.

Golijanin, D., Sherman, Y., Shapiro, A., and Pode, D. (1995). Detection of bladder tumors by immunostaining of the lewis x antigen in cells from voided urine. *Urology*, 46(2):173–177.

Guo, J. and Riebler, A. (2015). meta4diag: Bayesian bivariate meta-analysis of diagnostic test studies for routine practice. *ArXiv e-prints*.

Hamza, T. H., Arends, L. R., van Houwelingen, H. C., and Stijnen, T. (2009). Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology*, 9:73.

Hamza, T. H., van Houwelingen, H. C., and Stijnen, T. (2007). Random effects meta-analysis of proportions: the binomial distribution should be used to model the within-study variability. *Journal of Clinical Epidemiology*, 61(1):41–51. doi: 10.1016/j.jclinepi.2007.03.016.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., and Sterne, J. A. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8:239–251.

Higgins, J. P. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration.

Hillman, B. J., Frank, R. A., and Abraham, B. C. (2013). The medical imaging & technology alliance conference on research endpoints appropriate for medicare coverage of new PET radiopharmaceuticals. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 54(9):1675–1679.

Houssami, N., Irwig, L., Simpson, J. M., McKessar, M., Blome, S., and Noakes, J. (2003). Sydney breast imaging accuracy study: Comparative sensitivity and specificity of mammography and sonography in young women with symptoms. *AJR. American journal of roentgenology*, 180(4):935–940.

Hoyer, A. and Kuss, O. (2016). Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. *Statistical Methods in Medical Research*.

Independent UK Panel on Breast Cancer Screening (2012). The benefits and harms of breast cancer screening: an independent review. *Lancet*, 380(9855):1778–1786.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., O'Callaghan, C. J., Tu, D., Tebbutt, N. C., Simes, R. J., Chalchal, H., Shapiro, J. D., Robitaille, S., Price, T. J., Shepherd, L., Au, H.-J., Langer, C., Moore, M. J., and Zalcberg, J. R. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *The New England Journal of Medicine*, 359(17):1757–1765.

Karrasch, S., Linde, K., Rücker, G., Sommer, H., Karsch-Völk, M., Kleijnen, J., Jörres, R. A., and Schneider, A. (2016). Accuracy of FENO for diagnosing asthma: a systematic review. *Thorax*. doi: 10.1136/thoraxjnl-2016-208704.

Knottnerus, J. A. and Muris, J. W. (2003). Assessment of the accuracy of diagnostic tests: the cross-sectional study. *Journal of Clinical Epidemiology*, 56(11):1118–1128.

Kobberling, J., Trampisch, H., and Windeler, J. (1990). Memorandum for the evaluation of diagnostic measures. *Journal of Clinical Chemistry and Clinical Biochemistry. Zeitschrift Für Klinische Chemie Und Klinische Biochemie*, 28(12):873–879.

Lee, C. K., Lord, S. J., Coates, A. S., and Simes, R. J. (2009). Molecular biomarkers to individualise treatment: assessing the evidence. *The Medical Journal of Australia*, 190(11):631–636.

Lijmer, J. G. and Bossuyt, P. M. M. (2009). Various randomized designs can be used to evaluate medical tests. *Journal of Clinical Epidemiology*, 62(4):364–373.

Long, Q., Little, R. J., and Lin, X. H. (2008). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*, 103(482):474–484. Long, Qi Little, Roderick J. Lin, Xihong.

Lord, S., Irwig, L., and Simes, R. (2006). When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Annals of Internal Medicine*, 144(11):850–855.

Lu, B. and Gatsonis, C. (2013). Efficiency of study designs in diagnostic randomized clinical trials. *Statistics in Medicine*, 32(9):1451–1466.

Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology*, 57(9):925–932.

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., and Takwoingi, Y. (2016). *Handbook for DTA Reviews*. Cochrane Methods Screening and Diagnostic Tests. Chapter 10.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., and Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment*, 4(34):1–154.

Mandrekar, S. J. and Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 27(24):4027–4034.

Martínez-Camblor, P. (2014). Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research*. doi: 10.1177/0962280214537047.

McCaffery, K. J., Turner, R., Macaskill, P., Walter, S. D., Chan, S. F., and Irwig, L. (2011). Determining the impact of informed choice: Separating treatment effects from the effects of choice and selection in randomized trials. *Medical Decision Making*, 31(2):229–236.

Menten, J. and Lesaffre, E. (2015). A general framework for comparative bayesian meta-analysis of diagnostic studies. *Bmc Medical Research Methodology*, 15.

Merlin, T., Lehman, S., Hiller, J. E., and Ryan, P. (2013). The "linked evidence approach" to assess medical tests: a critical analysis. *International Journal of Technology Assessment in Health Care*, 29(3):343–350.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Moons, K. G., van Es, G. A., Michel, B. C., Büller, H. R., Habbema, J. D., and Grobbee, D. E. (1999). Redundancy of single diagnostic test evaluation. *Epidemiology (Cambridge, Mass.)*, 10(3):276–281.

Moses, L., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12(14):1293–1316.

MSAC (2005). *Guidelines for the assessment of diagnostic technologies*. Canberra, ACT: Commonwealth of Australia.

Mulherin, S. A. and Miller, W. C. (2002). Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, 137(7):598–602.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872.

Ng, S.-H., Chan, S.-C., Liao, C.-T., Chang, J. T.-C., Ko, S.-F., Wang, H.-M., Chin, S.-C., Lin, C.-Y., Huang, S.-F., and Yen, T.-C. (2008). Distant metastases and synchronous second primary tumors in patients with newly diagnosed oropharyngeal and hypopharyngeal carcinomas: evaluation of (18)f-FDG PET and extended-field multi-detector row CT. *Neuroradiology*, 50(11):969–979.

Nienhuis, S. J., Vles, J. S., Gerver, W. J., and Hoogland, H. J. (1997). Doppler ultrasonography in suspected intrauterine growth retardation: a randomized clinical trial. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 9(1):6–13.

Norlund, A., Marké, L.-A., af Geijerstam, J.-L., Oredsson, S., Britton, M., and OCTOPUS Study (2006). Immediate computed tomography or admission for observation after mild head injury: cost comparison in randomised controlled trial. *BMJ (Clinical research ed.)*, 333(7566):469.

Nyaga, V. N., Aerts, M., and Arbyn, M. (2016a). ANOVA model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research*. [Epub ahead of print].

Nyaga, V. N., Arbyn, M., and Aerts, M. (2016b). Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research*.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series 28.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M., and Yasui, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–1061.

Perucchini, D., Fischer, U., Spinas, G., Huch, R., Huch, A., and Lehmann, R. (1999). Using fasting plasma glucose concentrations to screen for gestational diabetes mellitus: prospective population based study. *British Medical Journal*, 319(7213):812–815.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., and Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, 31(4):463–479.

Poulsen, M. H., Bouchelouche, K., Høilund-Carlsen, P. F., Petersen, H., Gerke, O., Steffansen, S. I., Marcussen, N., Svolgaard, N., Vach, W., Geertsen, U., and Walter, S. (2012). [18f]fluoromethylcholine (FCH) positron emission tomography/computed tomography (PET/CT) for lymph node staging of prostate cancer: a prospective study of 210 patients. *BJU international*, 110(11):1666–1671.

Putter, H., Fiocco, M., and Stijnen, T. (2010). Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biometrical Journal*, 52(1):95–110.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reitsma, J., Glas, A., Rutjes, A., Scholten, R., Bossuyt, P., and Zwinderman, A. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990.

Riley, R. D., Ahmed, I., Ensor, J., Takwoingi, Y., Kirkham, A., Morris, R. K., Noordzij, J. P., and Deeks, J. J. (2015a). Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Systematic Reviews*, 4:12.

Riley, R. D., Elia, E. G., Malin, G., Hemming, K., and Price, M. P. (2015b). Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. *Statistics in Medicine*, 34(17):2481–2496.

Riley, R. D., Takwoingi, Y., Trikalinos, T., Guha, A., Biswas, A., Ensor, J., Morris, R. K., and Deeks, J. J. (2014). Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. *Journal of Biometrics and Biostatistics*, 5:196. 10.4172/2155-6180.1000196.

Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection, and treatment preference effects. *Statistics in Medicine*, 8:477–485.

Rücker, G. and Schumacher, M. (2009). Letter to the editor. *Biostatistics*, 10(4):806–807.

Rutgers, E., Piccart-Gebhart, M. J., Bogaerts, J., Delaloge, S., Veer, L. V. t., Rubio, I. T., Viale, G., Thompson, A. M., Passalacqua, R., Nitz, U., Vindevoghel, A., Pierga, J.-Y., Ravdin, P. M., Werutsky, G., and Cardoso, F. (2011). The EORTC 10041/BIG 03-04 MINDACT trial is feasible: results of the pilot phase. *European Journal of Cancer (Oxford, England: 1990)*, 47(18):2742–2749.

Rutjes, A. W. S., Reitsma, J. B., Vandenbroucke, J. P., Glas, A. S., and Bossuyt, P. M. M. (2005). Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry*, 51(8):1335–1341.

Rutter, C. M. and Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20:2865–2884.

Sackett, D. L. and Haynes, R. B. (2002). The architecture of diagnostic research. *BMJ*, 324(7336):539–541.

Sargent, D. J., Conley, B. A., Allegra, C., and Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 23(9):2020–2027.

Schaafsma, J. D., van der Graaf, Y., Rinkel, G. J. E., and Buskens, E. (2009). Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. *Journal of Clinical Epidemiology*, 62(12):1248–1252.

Schaefer, P. J., Boudghene, F. P., Brambs, H. J., Bret-Zurita, M., Caniego, J. L., Coulden, R. A., Gehl, H. B., Hammerstingl, R., Huber, A., Mendez, R. J., Nonent, M., Oestmann, J. W., Pueyo, J. C., Thurnher, S., Weishaupt, D., and Jahnke, T. (2006). Abdominal and iliac arterial stenoses: Comparative double-blinded randomized study of diagnostic accuracy of 3d MR angiography with gadodiamide or gadopentetate dimeglumine. *Radiology*, 238(3):827–840.

Schatzkin, A., Connor, R. J., Taylor, P. R., and Bunnag, B. (1987). Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. example of automated cytometry for early detection of cervical cancer. *American Journal of Epidemiology*, 125(4):672–678.

Scheibler, F., Raatz, H., Suter, K., Janssen, I., Grosselfinger, R., Schröer-Günther, M., Kunz, R., and Lange, S. (2010). [benefit assessment of PET in malignant lymphomas. the IQWiG point of view]. *Nuklearmedizin. Nuclear Medicine*, 49(1):1–5.

Scheibler, F., Zumbé, P., Janssen, I., Viebahn, M., Schröer-Günther, M., Grosselfinger, R., Hausner, E., Sauerland, S., and Lange, S. (2012). Randomized controlled trials on PET: a systematic review of topics, design, and quality. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 53(7):1016–1025.

Schröder, F., Hugosson, J., Roobol, M., Tammela, T., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., Denis, L., Recker, F., Berenguer, A., Määttänen, L., Bangma, C., Aus, G., Villers, A., Rebillard, X., van der Kwast, T., Blijenberg, B., Moss, S., de Koning, H., Auvinen, A., and ERSPC Investigators (2009). Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360(13):1320–1328. doi: 10.1056/NEJMoa0810084.

Schulz, K. F., Altman, D. G., Moher, D., and for the CONSORT Group (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340(mar23 1):c332–c332.

Schünemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Bossuyt, P., Chang, S., Muti, P., Jaeschke, R., and Guyatt, G. H. (2008a). GRADE: assessing the quality of evidence for diagnostic recommendations. *Evidence Based Medicine*, 13(6):162–163.

Schünemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G. E., Williams, J. W., Kunz, R., Craig, J., Montori, V. M., Bossuyt, P., and Guyatt, G. H. (2008b). Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*, 336(7653):1106–1110.

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1343.

Simon, R. (2010). Moving from correlative science to predictive oncology. *The EPMA journal*, 1(3):377–387.

Sox, H., Stern, S., Owens, D., and Abrams, H. L. (1989). *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions: Monograph of the Council on Health Care Technology*. National Academies Press (US), Washington (DC).

StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Steinhauser, S., Schumacher, M., and Rücker, G. (2016). Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Medical Research Methodology*, 16:97:97. DOI: 10.1186/s12874-016-0196-1.

Sutton, A. J., Cooper, N. J., Goodacre, S., and Stevenson, M. (2008). Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 28(5):650–667.

Takwoingi, Y., Leeflang, M. M. G., and Deeks, J. J. (2013). Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine*, 158(7):544–554.

Trikalinos, T. A., Hoaglin, D. C., Small, K. M., Terrin, N., and Schmid, C. (2014). Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods*, 5(4):294–312. doi: 10.1002/jrsm.1115.

Trikalinos, T. A., Siebert, U., and Lau, J. (2009). Decision-analytic modeling to evaluate benefits and harms of medical tests: Uses and limitations. *Medical Decision Making*, 29(5):E22–E29.

Vach, W., Gerke, O., and Høilund-Carlsen, P. F. (2012). Three principles to define the success of a diagnostic study could be identified. *Journal of Clinical Epidemiology*, 65(3):293–300.

Vach, W., Høilund-Carlsen, P. F., Gerke, O., and Weber, W. A. (2011). Generating evidence for clinical benefit of PET/CT in diagnosing cancer patients. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 52 Suppl 2:77S–85S.

van Houwelingen, H. C., Zwinderman, K. H., and Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, 12:2273–2284.

van Tinteren, H., Hoekstra, O. S., Smit, E. F., van den Bergh, J. H., Schreurs, A. J., Stallaert, R. A., van Velthoven, P. C., Comans, E. F., Diepenhorst, F. W., Verboom, P., van Mourlk, J. C., Postmus, P. E., Boers, M., and Teule, G. J. (2002). Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. *The Lancet*, 359(9315):1388–1392.

Vickers, A., Van Calster, B., and Steyerberg, E. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *British Medical Journal*, 352:i6. doi: 10.1136/bmj.i6.

Vickers, A., Van Calster, B., and Steyerberg, E. (2017). Decision curves, calibration, and subgroups. *Journal of Clinical Oncology*, 35(4):472–473. doi: 10.1200/JCO.2016.69.1576.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., and Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Annals of Internal Medicine*, 147(8):573–577.

Wason, J., Marshall, A., Dunn, J., Stein, R., and Stallard, N. (2014). Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *British Journal of Cancer*, 110(8):1950–1957. doi: 10.1038/bjc.2014.156.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., and Kleijnen, J. (2003). The development of quadas: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *Bmc Medical Research Methodology*, 3:25.

Willis, B. H. and Quigley, M. (2011). Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Medical Research Methodology*, 11:27.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Young, K. Y., Laird, A., and Zhou, X. H. (2010). The efficiency of clinical trial designs for predictive biomarker validation. *Clinical Trials*, 7(5):557–566.

Ziegler, A., Koch, A., Krockenberger, K., and Grosshennig, A. (2012). Personalized medicine using DNA biomarkers: a review. *Human Genetics*, 131(10):1627–1638.

Zwarenstein, M., Treweek, S., Gagnier, J. J., Altman, D. G., Tunis, S., Haynes, B., Oxman, A. D., Moher, D., and for the CONSORT and Pragmatic Trials in Healthcare (Practihc) groups (2008). Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ*, 337(nov11 2):a2390–a2390.