THE EVIDENCE BASE OF CLINICAL DIAGNOSIS

BMJ Books

THE EVIDENCE BASE OF CLINICAL DIAGNOSIS

THE EVIDENCE BASE OF CLINICAL DIAGNOSIS

Edited by

J ANDRÉ KNOTTNERUS
Netherlands School of Primary Care Research, University of Maastricht, The Netherlands

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First published in 2002
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 7279 1571 1

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed and bound in Spain by GraphyCems, Navarra

Preface

"I consider much less thinking has gone into the theory underlying diagnosis, or possibly one should say less energy has gone into constructing the correct model of diagnostic procedures, than into therapy or prevention where the concept of 'altering the natural history of the disease' has been generally accepted and a theory has been evolved for testing hypotheses concerning this."[1]

Although seeking an evidence base for medicine is as old as medicine itself, in the past decade the concept of evidence-based medicine (EBM) has strongly stimulated the application of the best available evidence from clinical research into medical practice. At the same time, this process has revealed the need for a more extensive and more valid evidence base as input for EBM. Accordingly, investigators have been encouraged to intensify the production and innovation of clinical knowledge, and clinical research has become more successful in seeing its results implemented in practice more completely in a shorter period.

In developing the evidence base of clinical management it has come forward that, even 3 decades after Archie Cochrane wrote the words cited above, the methodology of diagnostic research lags far behind that of research into the effectiveness of treatment. This is the more challenging because making an adequate diagnostic process is a prime requirement for appropriate clinical decision making, including prognostic assessment and the selection of the most effective treatment options.

In view of this apparent need for further methodological development of the evidence base of clinical diagnosis, this book was initiated. The aim is

Contents

Contributors

Jan H van Bemmel Department of Medical Informatics, Erasmus University Rotterdam, The Netherlands

Patrick M Bossuyt Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands

Frank Buntinx Department of General Practice, Catholic University Leuven, Belgium

Walter L Devillé Institute for Research in Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands

Geert-Jan Dinant Department of General Practice, University of Maastricht, The Netherlands

René Eijkemans Center for Clinical Decision Sciences, Department of Public Health, Erasmus University Rotterdam, The Netherlands

Arthur S Elstein Department of Medical Education, University of Illinois College of Medicine, Chicago, Illinois, USA

Constantine Gatsonis Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA

Paul P Glasziou Department of Social and Preventive Medicine, University of Queensland Medical School, Australia

J Dik F Habbema Center for Clinical Decision Sciences, Department of Public Health, Erasmus Medical Center Rotterdam, The Netherlands

R Brian Haynes Clinical Epidemiology and Biostatistics, McMaster University Medical Centre, Hamilton, Ontario, Canada

Les M Irwig Department of Public Health and Community Medicine, University of Sydney, Australia

J André Knottnerus Netherlands School of Primary Care Research, University of Maastricht, The Netherlands

Pieta Krijnen Center for Clinical Decision Sciences, Department of Public Health, Erasmus Medical Center Rotterdam, The Netherlands

Johan van der Lei Department of Medical Informatics, Erasmus University Rotterdam, The Netherlands

Jeroen G Lijmer Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands

Jean W Muris Department of General Practice, University of Maastricht, The Netherlands

David L Sackett Trout Research and Education Centre at Irish Lake, Markdale, Ontario, Canada

Onno P van Schayck Institute for Extramural and Transmural Health Care, University of Maastricht, The Netherlands

Alan Schwartz Department of Medical Education, University of Illinois College of Medicine, Chicago, Illinois, USA

Chris van Weel Department of General Practice and Social Medicine, Institute for Evidence-Based Practice, University of Nijmegen, The Netherlands

Ron AG Winkens Transmural and Diagnostic Centre, Academic Hospital Maastricht, The Netherlands

to provide a comprehensive framework for (future) investigators who want to do diagnostic research, and for clinicians, practitioners and students who are interested to learn more about principles, and about relevant methodological options and pitfalls. Clearly, not all topics relevant for diagnostic research could be covered, nor could the selected subjects be dealt with in all detail. For those who wish to know more, the references in the chapters can be a useful guide. In preparing the work, the contributors were able to profit from the experience and insights collected and reported by many leading clinical researchers in the field.

First, a general outline of diagnostic research is presented. What are the key objectives, the challenges, and the corresponding options for study design? What should the architecture of diagnostic research look like to provide us with an appropriate research strategy, yielding the clinical information we are looking for, with a minimum burden for study patients and an efficient use of resources? Second, important design features for studying the accuracy and clinical impact of diagnostic tests and procedures are dealt with in more detail, addressing the cross-sectional study, the randomised trial, and the before–after study. In addition, it is shown that the impact of diagnostic tests varies with different clinical settings and target populations, and indications are given as how to ensure that estimates of test accuracy will travel and be transferable to other settings. Also, for clinical diagnostic studies, an overview of the most important data-analytic issues is presented, from simple two by two tables to multiple logistic regression analysis.

Nowadays, for both clinical investigators and readers of research articles, it is not enough to understand the methodology of original clinical studies. They must also know more about the techniques to summarise and synthesise results from various clinical studies on a similar topic. Guidelines for diagnostic systematic reviews and meta-analysis are therefore presented.

Learning from accumulated clinical experience and the application of diagnostic knowledge in practice has much in common with retrieving and selecting information from clinical databases. Accordingly, diagnostic decision support using information and communication technology (ICT) is addressed as an increasingly important domain for clinical practice, research, and education. Furthermore, as clinical research results can only be successfully incorporated into diagnostic decision making if the way clinicians tend to solve medical problems is taken into account, an overview of the domain of clinical problem solving is given. Eventually, we have to recognise that improving test use in daily care needs more than clinical research, and presenting guidelines and other supportive materials. Therefore, the strategy of successful implementation – which has become a field of study in itself – is also covered.

This book includes contributions from many authors. In order to allow each chapter to keep a logical structure in itself, a number of topics have

been dealt with more than once, albeit to a varying extent. Instead of seeing this as a problem, we think that it may be informative for readers to see important issues considered from different perspectives.

Parallel to the preparation of this book, an initiative to reach international agreement of standards for reporting diagnostic accuracy (STARD) was taken and elaborated. This development can be expected to make a significant contribution to improving the quality of published literature on the value of diagnostic tests. We are happy that members of the group of initiators of STARD have contributed to this book as the authors of chapters 4 and 6, and are looking forward to seeing these standards having an impact.

The field of diagnostic research is developing strongly and an increasing number of talented clinical investigators are working in (the methodology of) diagnostic research. In view of this dynamic field, we welcome comments from readers and suggestions for possible improvements.

The contributors wish to thank Richard Smith from the *BMJ*, who has so positively welcomed the initiative for this book, Trish Groves from the *BMJ* who gave very useful feedback on the proposed outline and stimulated us to work it out, and Mary Banks from BMJ Books, who has provided support and encouragement from the beginning and monitored the progress of the book until the work was done.

André Knottnerus

1 Cochrane AL. Effectiveness and efficiency. Random reflections on health services. The Nuffield Provincial Hospitals Trusts, 1972. Reprinted: London, the Royal Society of Medicine Press Limited, 1999.

1 General introduction: evaluation of diagnostic procedures

J ANDRÉ KNOTTNERUS, CHRIS VAN WEEL

Summary box

- Whereas the development of diagnostic technologies has greatly accelerated, the methodology of diagnostic research lags far behind that of evaluation of treatment.
- Objectives of diagnostic testing are (1) detecting or excluding disorders, (2) contributing to further diagnostic and therapeutic management, (3) assessing prognosis, (4) monitoring clinical course, and (5) measuring general health or fitness.
- Methodological challenges include dealing with complex relations, the "gold standard" problem, spectrum and selection bias, "soft" outcome measures, observer variability and bias, addressing clinical relevance, appropriate sample size, and rapid progress of applicable knowledge over time.
- Choosing the appropriate study design depends on the research question; the most important designs are the cross-sectional study (to determine the accuracy and added value of diagnostic procedures) and the randomised controlled trial (to evaluate the clinical impact of testing).
- In order to synthesise the results of various studies on the same topic, diagnostic systematic reviews and meta-analyses are powerful tools.
- To make the step from research to practice, clinical decision analysis, cost effectiveness studies, and quality of care research, including implementation studies, are indispensible.

Introduction

The development and introduction of new diagnostic technologies have accelerated greatly over the past few decades. This is reflected in a substantial expansion of research on diagnostic tests. For example, the number of such publications we found in MEDLINE increased from about 2000 in the period 1966–1970 to about 17 000 in the period 1996–2000. However, the evaluation of diagnostic techniques is far from being as advanced as the evaluation of therapies.

At present, unlike the situation with regard to drugs, there are no formal requirements that a diagnostic test must meet in order to be accepted or retained as a routine part of health care. This is related to another point: in spite of useful early initiatives[1,2] the methodology for evaluation of diagnostics is not much crystallised, in contrast to the deeply rooted consensus regarding the principles of the randomised controlled trial on therapeutic effectiveness[1,3] and the broad agreement on aetiologic study designs.[4,5] It is not surprising, then, that serious methodological flaws are often found in published diagnostic studies.[6–8] A further point of concern is that the funding of diagnostic evaluation studies is poorly organised, especially if the research is not focused on particular body systems or categories of disorders well covered by research foundations. Rather than being limited to a particular body system, diagnostic evaluation studies frequently start from a complaint, a clinical problem, or certain tests.

The first crucial medical intervention in an episode of illness is diagnostic, labelling symptoms and complaints as illness, and indicating possible disease and its prognosis. Effective and efficient therapy – including reassurance, "watchful waiting" and supporting patient self-efficacy – depends to a large extent on an accurate interpretation of (early) symptoms and the outcome of the diagnostic process. Therefore, because the quality of diagnostic procedures is indicative for the quality of health care as a whole, it is vital to overcome the shortfall in standards, methodology, and funding. Accurate evaluation of diagnostic performance will contribute to the prevention of unjustified treatment, lack of treatment or mistreatment, as well as unnecessary costs.

This introductory chapter presents an overview of the objectives of diagnostic testing and evaluation research, important methodological challenges, and research design options.

Objectives

Diagnostic testing can be seen as the collection of additional information with the intention of (further) clarifying the character and prognosis of the patient's condition, and can include patients' characteristics, symptoms and signs, history and physical examination items, or additional tests using

laboratory or other technical facilities. Not only a "test" must be considered, but also the specific question the test is supposed to answer. Therefore, the performance of tests must be evaluated in accordance with their intended objectives. Objectives may include:

- *Detecting or excluding disorders, by increasing diagnostic certainty as to their presence or absence*. This can only be achieved if the test has sufficient discrimination. Table 1.1 shows the most common measures of discrimination. Most of these can be simply derived from a 2×2 table comparing the test result with the diagnostic standard, as demonstrated by the example of ankle trauma. A more elaborate and comprehensive explanation of how to calculate these and other measures from collected data is presented in Chapter 7. Examples of tests for which such measures have been assessed are given in Table 1.2. Such a representation allows various tests for the same purpose to be compared. This can show, for example, that less invasive tests (such as ultrasonography) may be as good as or even better diagnostically than more invasive or hazardous ones (for example angiography). Also, it can be shown that history data (for example change in bowel habit) may be at least as valuable as laboratory data. What is important is not just the discrimination per se, but rather what a test may add to what cheaper and less invasive diagnostics already provide to the diagnostic process. This is relevant, for instance, in assessing the added value of liver function tests to history taking and physical examination in ill-defined, non-specific complaints.
- *Contributing to the decision making process with regard to further diagnostic and therapeutic management*, including the indications for therapy (for example by determining the localisation and shape of a lesion) and choosing the preferred therapeutic approach
- *Assessing prognosis* on the basis of the nature and severity of diagnostic findings. This is a starting point for planning the clinical follow up and for informing and – if justified – reassuring the patient
- *Monitoring the clinical course* of a disorder or a state of health such as pregnancy, or the clinical course of an illness during or after treatment
- *Measuring physical fitness* in relation to requirements, for example for sports or employment.

The evaluation of a diagnostic test concentrates on its added value for the intended application, taking into consideration the burden for the patient and any possible complications resulting from the test (such as intestinal perforation in endoscopy). This requires a comparison between the situations with and without the use of the test, or a comparison with the use of other tests.

Prior to evaluation, one must decide whether to focus on maximising the health perspectives of the individual patient (which is usually the

Table 1.1 Commonly used measures of the discrimination of a diagnostic test T for disease D, illustrated with physical examination for detecting a fracture in ankle trauma, using *x* ray film as the reference standard.

	D: result of *x* ray		
T: conclusion of physical examination	Fracture	No fracture	Total
Fracture	190	80	270
No fracture	10	720	730
Total	200	800	1000

The SENSITIVITY of T is the probability of a positive test result in people with D: $P(T+|D+) = 190/200 = 0.95$.
The SPECIFICITY of T is the probability of a negative test result in people without D: $P(T-|D-) = 720/800 = 0.90$.

Note: sensitivity and specificity together determine the discrimination of a test.

The LIKELIHOOD RATIO (LR) of test result T_X is the probability of test result T_X in people with D, divided by the probability of T_X in people without D.

The general formula for LR_X is: $\dfrac{P(T_X|D+)}{P(T_X|D-)}$

For a positive result, LR+ is: $\dfrac{P(T+|D+)}{P(T+|D-)}$

which is equivalent to: $\dfrac{\text{Sensitivity}}{1-\text{specificity}} = \dfrac{190/200}{1-720/800} = 9.5$

For a negative result, LR− is: $\dfrac{P(T-|D+)}{P(T-|D-)}$

which is equivalent to: $\dfrac{1-\text{sensitivity}}{\text{Specificity}} = \dfrac{1-190/200}{720/800} = 0.06$

Note: LR is an overall measure of the discrimination of test result T_X. The test is useless if LR = 1. The test is better the more LR differs from 1, that is, greater than 1 for LR+ and lower than 1 for LR−.

For tests with multiple outcome categories, LR_X can be calculated for every separate category x as the ratio of the probability of outcome category x among diseased and the probability of outcome category x among non-diseased.

The PREDICTIVE VALUE of a test result T_X is:
for a positive result, the probability of D in persons with a positive test result: $P(D+|T+) = 190/270 = 0.70$.
for a negative result, the probability of absence of D in persons with a negative result: $P(D-|T-) = 720/730 = 0.99$.

Note: the predictive value (posterior or post-test probability) must be compared with the estimated probability of D before T is carried out (the prior or pretest probability). For a good discrimination, the difference between the post-test and the pretest probability should be large.

The ODDS RATIO (OR), or the cross-product ratio, represents the overall discrimination of a dichotomous test T, and is equivalent to the ratio of LR+ and LR−. $OR = (190 \times 270)/(80 \times 10) = 171$

Note: If OR = 1, T is useless. T is better the more OR differs from 1.

The RECEIVER OPERATING CHARACTERISTIC (ROC) curve represents the relation between sensitivity and specificity for tests with a variable cut-off point, on an ordinal scale (for example, in case of 5 degrees of suspicion of ankle fracture; or cervical smear) or interval scale (for example, if degree of suspicion of ankle fracture is expressed in a percentage; or ST changes in exercise ECG testing). If the AUC (area under the curve) = 0.5, the test is useless. For a perfect test the AUC = 1.0 (see Chapter 7).

Table 1.2 Discrimination of some diagnostic tests for various target disorders, expressed in sensitivity, specificity, likelihood ratios, and odds ratio (estimates based on several sources).

Test	Target Disorder	Sensitivity (%)	Specificity (%)	Likelihood ratio		Odds ratio
				Positive result	Negative result	
Exercise ECG*[9]	Coronary stenosis	65	89	5.9	0.39	15.0
Stress thallium scintigraphy[9]	Coronary stenosis	85	85	5.7	0.18	32.1
Ultrasonography[9]	Pancreatic cancer	70	85	4.7	0.35	13.2
CT scan[9]	Pancreatic cancer	85	90	8.5	0.17	51.0
Angiography[9]	Pancreatic cancer	75	80	3.8	0.31	12.0
ESR ⩾28 mm/1 h**[10]	Malignancy	78	94	13.0	0.23	56.0
ESR ⩾28 mm/1 h**[10]	Inflammatory disease	46	95	9.2	0.57	16.2
Intermittent claudication**[11]	Peripheral arterial occlusive disease	31	93	4.4	0.74	5.6
Posterior tibial/dorsalis pedis artery pulse**[11]	Peripheral arterial occlusive disease	73	92	9.1	0.29	30.4
Change in bowel habit**[12]	Colorectal cancer	88	72	3.1	0.17	18.4
Weight loss**[12]	Colorectal cancer	44	85	2.9	0.66	4.6
ESR ⩾30 mm/1 h**[12]	Colorectal cancer	40	96	10.0	0.42	14.0
White blood cell count $>10^9$**[12]	Colorectal cancer	75	90	7.5	0.28	26.3
Occult blood test ⩾1 positive out of 3**[12]	Colorectal cancer	50	82	2.7	0.61	4.6

*Cut-off point: ST depression ⩾1 mm.
**In a general practice setting.

physician's aim) or on the best possible cost effectiveness (as economists are likely to do). The latter can be expressed in the amount of money to be invested per number of life years gained, whether or not adjusted for quality of life. Between these two approaches, which do not necessarily yield the same outcome, there is the tension between strictly individual and collective interests. This becomes especially obvious when policy makers have to decide which options would be accepted as the most efficient in a macroeconomic perspective.

Another prior decision is whether one would be satisfied with a qualitative understanding of the diagnostic decision making process, or is also aiming at a detailed quantitative analysis.[13] In the first case one would chart the stages

and structure of the decision making process, in relation to the test to be evaluated. This may already provide sufficient insight, for instance if it becomes clear beforehand that the result will not influence the decision to be taken. Examples of useless testing are (1) the value of the routine electrocardiogram in acute chest pain, exploring the likelihood of a suspected myocardial infarction, with the consequent decision whether or not to admit the patient to hospital; and (2) the value of "routine blood tests" in general practice for the decision as to whether or not to refer a patient with acute abdominal pain to a surgeon. In addition to qualitatively mapping the structure of the decision making process, quantitative analysis attempts to assess test discrimination and the ultimate clinical outcome, taking the risks (and the costs) of the test procedure into account. The choice of a qualitative or a quantitative approach depends on the question to be answered and the data available.

If a test has not yet been introduced, the prospects for a good evaluation are better than if it is already in general use. It is then, for example, still possible to define an appropriate control group to whom the test is not applied, so that its influence on the prognosis can be investigated. In addition, at such an early stage the conclusion of the analysis can still be used in the decision regarding introduction. Furthermore, it is possible to plan a procedure for monitoring and evaluation after introduction. All of this emphasises the importance of developing an evaluation programme before a test is introduced.

A common misunderstanding is that only expensive, advanced diagnostic technology cause unacceptable increases in healthcare costs; in fact, cheap but very frequently used (routine) tests account for a major part of direct costs. Moreover, these tests greatly influence other costs, as they often preselect patients for more expensive procedures. Yet the performance of such low-threshold diagnostics has often not been adequately evaluated. Examples include many applications of haematological, clinicochemical, and urine tests.[14–16]

Methodological challenges

In the evaluation of diagnostic procedures a number of methodological challenges have to be considered.

Complex relations

Most diagnostics have more than one indication or are relevant for more than one nosological outcome. In addition, tests are often not applied in isolation but in combinations, for instance in the context of protocols. Ideally, diagnostic research should reflect the healthcare context,[17] but it is generally impossible to investigate all aspects in one study. Therefore,

choices must be made as to which issues are the most important. Multivariable statistical techniques are available to allow for the (added) value of various diagnostic data, both separately and in combination, and also in the form of diagnostic prediction rules.[18,19] Such techniques were originally developed for the purpose of analysing aetiologic data, generally focusing on the overall aetiologic impact of a factor adjusted for covariables. Diagnostic analysis aims to specify test performance in clinical subgroups or to identify the set of variables that yield the best individual diagnostic prediction, which is a completely different perspective. Much work remains to be done to improve the methodology of diagnostic data analysis.[20]

Diagnostic data analysis will be discussed further in Chapter 7.

The "gold" standard problem

To evaluate the discriminatory power of a test, its results must be compared with an independently established standard diagnosis. However, a "gold" standard, providing full certainty on the health status, rarely exists. Even *x* rays, CT scans and pathological preparations may produce false positive and false negative results. The aim must then be to define an adequate reference standard that approximates the "gold" standard as closely as possible.

Sometimes one is faced with the question whether any appropriate reference standard procedure exists at all. For example, in determining the discrimination of liver tests for diagnosing liver pathology, neither imaging techniques nor biopsies can detect all abnormalities. In addition, as a liver biopsy is an invasive procedure it is unsuitable for use as a standard in an evaluation study. A useful independent standard diagnosis may not even exist conceptually, for example when determining the predictive value of symptoms that are themselves part of the disease definition, as in migraine, or when the symptoms and functionality are more important for management decisions than the anatomical status, as in prostatism. Also, in studying the diagnostic value of clinical examination to detect severe pathology in non-acute abdominal complaints, a comprehensive invasive standard diagnostic screening, if at all possible or ethically allowed, would yield many irrelevant findings and not all relevant pathology would be immediately found. An option, then, is diagnostic assessment after a follow up period by an independent panel of experts, representing a "delayed type" cross-sectional study.[21] This may not be perfect, but can be the most acceptable solution.[1]

A further issue is the dominance of prevailing reference standards. For example, as long as classic angiography is considered the standard when validating new vascular imaging techniques, the latter will always seem less valid because perfect agreement is never attainable. However, as soon as the new method comes to be regarded as sufficiently valid to be accepted as the

standard, the difference will from then on be explained in favour of this new method. In addition, when comparing advanced ultrasound measurements in blood vessels with angiography, one must accept that the two methods actually measure different concepts: the first measures blood flow, relevant to explain the symptoms clinically, whereas the second reflects the anatomical situation, which is important for the surgeon. Furthermore, the progress of clinicopathological insights is of great importance. For example, although clinical pattern X may first be the standard to evaluate the significance of microbiological findings, it will become of secondary diagnostic importance once the infectious agent causing X has been identified. The agent will then be the diagnostic standard, as illustrated by the history of the diagnosis of tuberculosis.

In Chapters 3 and 6 more will be said about reference standard problems.

Spectrum and selection bias

The evaluation of diagnostics may be flawed by many types of bias.[1,22,23] The most important of these are spectrum bias and selection bias.

Spectrum bias may occur when the discrimination of the diagnostic is assessed in a study population with a different clinical spectrum (for instance in more advanced cases) than will be found among those in whom the test is to be applied in practice. This may, for example, happen with tests calibrated in a hospital setting but applied in general practice. Also, sensitivity may be determined in seriously diseased subjects, whereas specificity is tested in clearly healthy subjects. Both will then be grossly overestimated relative to the practical situation, where testing is really necessary because it is clinically impossible to distinguish in advance who is healthy and who is diseased.

Selection bias is to be expected if there is a relation between the test result and the probability of being included in the study population in which the test is calibrated. For example, subjects with an abnormal exercise electrocardiogram are relatively likely to be preselected for coronary angiography. Consequently, if this exercise test is calibrated among preselected subjects, a higher sensitivity and a lower specificity will be found than if this preselection had not occurred.[24] Similarly, on the basis of referral patterns alone it is to be expected that the sensitivity of many tests is higher in the clinic than in general practice, and the specificity lower.

Although spectrum and selection biases are often related, in the first the clinical picture is the primary point of concern, whereas in the latter the mechanism of selection is the principal issue. These types of bias may affect not only sensitivity and specificity, but also all other measures of discrimination listed in Table 1.1.[25]

Chapters 2 and 6 will further address the issue of dealing with spectrum and selection biases.

"Soft" measures

Subjective factors such as pain, feeling unwell and the need for reassurance are of great importance in diagnostic management. Most decisions for a watchful waiting strategy in the early phase of an episode of illness are based on the valuation of "soft" measures. These often determine the indication for diagnostic examinations, and may themselves be part of the diagnostics (for example a symptom or complaint) to be evaluated. Also, such factors are generally indispensable in the assessment of the overall clinical outcome. Evaluation studies should, on the one hand, aim as much as possible to objectify these subjective factors in a reproducible way. On the other hand, interindividual and even intraindividual differences will always play a part[26] and should be acknowledged in the clinical decision making process.

Observer variability and observer bias

Variability between different observers, as well as for the same observer in reading and interpreting diagnostic data, should not only be acknowledged for "soft" diagnostics such as history taking and physical examination, but also for "harder" ones like *x* rays, CT scans and pathological slides. Even tests not involving any human factors show inter- and intrainstrument variability. Such variability should be limited if the diagnostic is to produce useful information.

At the same time, evaluation studies should beware of systematic observer bias as a result of prior knowledge about the subjects examined. Clearly, if one wishes to evaluate whether a doctor can accurately diagnose an ankle fracture based on history and clinical examination, it must be certain that he is not aware of an available *x* ray result; and a pathologist making an independent final diagnosis should not be informed about the most likely clinical diagnosis.[27] In such situations "blinding" is required. A different form of observer bias could occur if the diagnosticians are prejudiced in favour of one of the methods to be compared, as they may unconsciously put greater effort into that technique. A further challenge is that the experience and skill required should be equal for the methods compared, if these are to have a fair chance in the assessment. In this respect, new methods are at risk of being disadvantaged, especially shortly after being introduced.

Discrimination does not mean usefulness

For various reasons, a test with very good discrimination does not necessarily influence management.

To begin with, a test may add too little to what is already known clinically to alter management. Furthermore, the physician may take insufficient

account of the information provided by the test. This is a complex problem. For instance, studies of the consequences of routine blood testing have shown that in some cases an unaltered diagnosis still led to changes in the considered policy.[14] In a study on the therapeutic impact of upper gastrointestinal endoscopy, a number of changes (23%) in management were made in the absence of a change in diagnosis, whereas in many patients (30%) in whom the diagnosis was changed management was not altered.[28] Also, a test may detect a disorder for which no effective treatment is available. For example, the MRI scan provides refined diagnostic information with regard to various brain conditions for which no therapy is yet in prospect. Finally, as discussed in the previous section, supplementary test results may not be relevant for treatment decisions.

For this reason we strongly recommend that evaluation studies investigate both the discrimination of a test and its influence on management.

Indication area and prior probability

Whether a test can effectively detect or exclude a particular disorder is influenced by the prior probability of that disorder. A test is generally not useful if the prior probability is either very low or very high: not only will the result rarely influence patient management, but the risk of, respectively, a false positive or a false negative result is relatively high. In other words, there is an "indication area" for the test between these extremes of prior probability.[9,10] Evaluation of diagnostics should therefore address the issue of whether the test could be particularly useful for certain categories of prior probability. For example, tests with a moderate specificity are not useful for screening in an asymptomatic population (with a low prior probability) because of the high risk of false positive results.

Small steps and large numbers

Compared with therapeutic effectiveness studies, evaluation studies of diagnostic procedures have often neglected the question of whether the sample size is adequate to provide the desired information with a sufficient degree of certainty. A problem is that progress in diagnostic decision making often takes the form of a series of small steps so as to gain in certainty, rather than one big breakthrough. Evaluating the importance of a small step, however, requires a relatively large study population.

Changes over time and the mosaic of evidence

Innovations in diagnostic technology may proceed at such a speed that a thorough evaluation may take longer than the development of even more advanced techniques. For example, the results of evaluation studies on the cost effectiveness of the CT scan had not yet crystallised when the MRI and PET scans appeared on the scene. So, the results of evaluation studies

may already be lagging behind when they appear. Therefore, there is a need for general models (scenarios) for the evaluation of particular (types of) tests and test procedures, whose overall framework is relatively stable and into which information on new tests can be entered by substituting the relevant piece in the whole mosaic. This allows, for instance, a quick evaluation of the impact of new mammographic or DNA techniques with better discrimination on the cost effectiveness of breast cancer screening, if other pieces of the mosaic (such as treatment efficacy) have not changed. As discrimination itself can often be relatively rapidly assessed by means of a cross-sectional study, this may avoid new prospective studies. The same can be said for the influence of changes in relevant costs, such as fees for medical treatment or the price of drugs.

Research designs

There are various methodological approaches for evaluating diagnostic technologies, including original clinical research on the one hand, and systematically synthesising the findings of already performed empirical studies and clinical expertise on the other.

For empirical clinical studies, a range of design options is available. The appropriate study design depends on the research question to be answered (Table 1.3). In diagnostic accuracy studies the relationship between test result and reference standard has to be assessed cross-sectionally. This can be achieved by a cross-sectional survey, but especially in early validation studies other approaches (case–referent or test result-based sampling) can be most efficient. Design options for studying the impact of diagnostic testing on clinical decision making and patient prognosis are the "diagnostic randomised controlled trial" (RCT), which is methodologically the strongest approach, and the before–after study. Also, cohort and case–control designs have been shown to have a place in this context. In Chapter 2, the most important strategic considerations in choosing the appropriate design in diagnostic research will be specifically addressed.

Current knowledge can be synthesised by systematic reviews, meta-analyses, clinical decision analysis, cost effectiveness studies and consensus methods, with the ultimate aim of integrating and translating research findings for implementation in practice.

In the following, issues of special relevance to diagnostic evaluation studies will be briefly outlined.

Clinical studies

A common type of research is the cross-sectional study, assessing the relationship between diagnostic test results and the presence of particular

Table 1.3 Methodological options in diagnostic research in relation to study objectives.

Study objective	Methodological options
Clinical studies	
Diagnostic accuracy	Cross-sectional study survey case–referent sampling test result-based sampling
Impact of diagnostic testing on prognosis or management	Randomised controlled trial Cohort study Case–control study Before–after study
Synthesising findings and expertise	
Synthesising results of multiple studies	Systematic review Meta-analysis
Evaluation of most effective or cost effective diagnostic strategy	Clinical decision analysis Cost effectiveness analysis
Translating findings for practice	Integrating results of the above mentioned approaches Expert consensus methods Developing guidelines
Integrating information in clinical practice	
	ICT support studies Studying diagnostic problem solving Evaluation of implementation in practice

disorders. This relationship is usually expressed in the measures of discrimination included in Table 1.1. Design options are: (1) a survey in an "indicated population", representing subjects in whom the studied test would be considered in practice; (2) sampling groups with (cases) and without disease (referents) to compare their test distributions; or (3) sampling groups with different test results, between which the occurrence of a disease is compared. It is advisable to include in the evaluation already adopted tests, as this is a direct way to obtain an estimate of the added value of the new test. The cross-sectional study will be dealt with in more detail in Chapter 3.

In an RCT the experimental group undergoes the test to be evaluated, while a control group undergoes a different (for example the usual) or no test. This allows the assessment of not only differences in the percentage of correct diagnoses, but also the influence of the evaluated test on management and prognosis. A variant is to apply the diagnostic test to all patients but to disclose its results to the caregivers for a random half of the patients, if ethically justified. This constitutes an ideal placebo procedure for the patient. Although diagnostic RCTs are not easy to carry out and often not feasible, several have been already carried out some time

ago.[29–34] Among the best known are the early trials on the effectiveness of breast cancer screening, which have often linked a standardised management protocol to the screening result.[35,36] The randomised controlled trial in diagnostic research is further discussed in Chapter 4.

If the prognostic value of a test is to be assessed and an RCT is not feasible, its principles can serve as the paradigm in applying other methods, one such being the cohort study. The difference from the RCT is that the diagnostic information is not randomly assigned, but a comparison is made between two previously established groups.[37] It has the methodological problem that one can never be sure, especially regarding unknown or unmeasurable covariables, whether the compared groups have similar disease or prognostic spectra to begin with. A method providing relatively rapid results regarding the clinical impact of a test is the case–control study. This is often carried out retrospectively, that is, after the course and the final status of the patients are known, in subjects who at the time have been eligible for the diagnostic test to be evaluated. It can be studied whether "indicated subjects" showing an adverse outcome (cases) underwent the diagnostic test more or less frequently than indicated subjects without such outcome (controls). A basic requirement is that the diagnostic must have been available to all involved at the time. Well known examples are case–control studies on the relationship between mortality from breast cancer and participation in breast cancer screening programmes.[38,39] This is an efficient approach, although potential bias because of lack of prior comparability of tested and non-tested subjects must once again be borne in mind.

The influence of a diagnostic examination on the physician's management can be also investigated by comparing the intended management policies before and after test results are available. Such before–after comparisons (diagnostic impact studies) have their own applications, limitations and precautionary measures, as reviewed by Guyatt et al.[40] The method has, for example, been applied in determining the added value of the CT scan and in studying the diagnostic impact of haematological tests in general practice.[41,42] The before–after study design will be outlined in Chapter 5.

Although using appropriate inclusion and exclusion criteria for study subjects is as important as in therapeutic research, in diagnostic research defining such criteria is less well developed. However, appropriate criteria are indispensible in order to focus on the clinical question at issue, the relevant spectrum of clinical severity, the disorders to be evaluated and the desired degree of selection of the study population (for example primary care or referred population).[43]

Synthesising research findings and clinical expertise

Often the problem is not so much a lack of research findings but the lack of a good summary and systematic processing of those findings.

A diagnostic systematic review, and meta-analysis of the pooled data of a number of diagnostic studies, can synthesise the results of those studies. This provides an overall assessment of the value of diagnostic procedures,[44,45] and can also help to identify differences in test accuracy between clinical subgroups. In this way, an overview of the current state of knowledge is obtained within a relatively short time. At present, this method is very much under development. One is aiming at bridging a methodological backlog, compared to the more established therapeutic systematic review and meta-analysis. The methodology of systematically reviewing studies on the accuracy of diagnostic tests is elaborated in Chapter 8.

Another important approach is clinical decision analysis, systematically comparing various diagnostic strategies as to their clinical outcome or cost effectiveness, supported by probability and decision trees. If good estimates of the discrimination and risks of testing, the occurrence and prognosis of suspected disorders, and the "value" of various clinical outcomes are available, a decision tree can be evaluated quantitatively in order to identify the clinically optimal or most cost effective strategy. An important element in the decision analytic approach is the combined analysis of diagnostic and therapeutic effectivenes. In this context, a qualitative analysis can be very useful. For example, non-invasive techniques nowadays show a high level of discrimination in diagnosing carotid stenoses, even in asymptomatic patients. This allows improved patient selection for the invasive and more hazardous carotid angiography, which is needed to make final decisions regarding surgical intervention. But if surgery has not been proved to influence the prognosis of asymptomatic patients favourably compared to non-surgical management,[46] the decision tree is greatly simplified as it no longer would include either angiography or surgery, and maybe not even non-invasive testing.

Decision analysis does not always provide an answer. The problem may be too complex to be summarised in a tree, essential data may be missing, and there is often a lack of agreement on key assumptions regarding the value of outcomes. Therefore, consensus procedures are often an indispensable step in the translational process from clinical research to guidelines for practice. In these procedures, clinical experts integrate the most recent state of knowledge with their experience to reach agreement on clinical guidelines regarding the preferred diagnostic approach of a particular medical problem, differentiated for relevant subgroups.[47,48]

Integrating information in clinical practice

In order to help clinical investigators harvest essential diagnostic research data from clinical databases and to support clinicians in making and

improving diagnostic decisions, medical informatics and ICT (information and communications technology) innovations are indispensible. However, as described in Chapter 9, to use the potentials in this field optimally, specific methodological and practical requirements must be met.

The information processing approaches outlined in the previous section constitute links between research findings and clinical practice, and can be applied in combination to support evidence-based medicine. How such input can have optimal impact on the diagnostic decision making of individual doctors is, however, far from simple or straightforward. Therefore, given the growing cognitive requirements of diagnostic techniques, studies to increase our insight in diagnostic problem solving by clinicians is an increasingly important part of diagnostic research. This topic is discussed in Chapter 10.

Information from good clinical studies, systematic reviews and guideline construction is necessary but in many cases not sufficient for improving routine practice. In view of this, during the last decade, implementation research has been strongly developed to face this challenge and to facilitate the steps from clinical science to patient care. Accordingly, Chapter 11 deals with implementation of (cost-)effective test ordering in clinical practice.

Conclusion

Diagnostic technology assessment would be greatly stimulated if formal standards for the evaluation of diagnostics were to be formulated, as a requirement for market acceptance. Health authorities could take the initiative in assembling panels of experts to promote and monitor the evaluation of both new and traditional diagnostic facilities. Criteria for the acceptance and retention of diagnostics in clinical practice should be developed. Furthermore, professional organisations have a great responsibility to set, implement, maintain, and improve clinical standards. More effective international cooperation would be useful, as it has proved to be in the approval and quality control of drugs. In this way, the availability of resources for industrial, private, and governmental funding for diagnostic technology assessment would also be stimulated.

As regards the feasibility of diagnostic evaluation studies, the required size and duration must be considered in relation to the speed of technological progress. This speed can be very great, for instance in areas where the progress of molecular genetic knowledge and information and communication technology play an important part. Especially in such areas, updating of decision analyses, expert assessments and scenarios by inserting new pieces of the "mosaic" of evidence may be more useful than fully comprehensive, lengthy trials. This may, for example, be very relevant for the evaluation of diagnostic areas where traditional tests will be replaced by

DNA diagnostics in the years to come. Finally, successful integration of "soft" health measures, quality of life aspects, and health economic objectives into clinical evaluations will require much additional research, and methodological and ethical consideration.[49]

References

1 Feinstein AR. *Clinical epidemiology. The architecture of clinical research*. Philadelphia: WB Saunders, 1985.
2 Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown and Co., 1985.
3 Pocock SJ. *Clinical trials, a practical approach*. New York: John Wiley & Sons, 1983.
4 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research, principles and quantitative methods*. Belmont (CA): Wadsworth, 1982.
5 Miettinen OS. *Theoretical epidemiology, principles of occurrence research in medicine*. New York: John Wiley & Sons, 1985.
6 Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA* 1984;**252**:2418–22.
7 Reid ML, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic research. Getting better but still not good. *JAMA* 1995;**274**:645–51.
8 Lijmer JG, Mol BW, Heisterkamp S, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
9 Panzer RJ, Black ER, Griner PF, eds. *Diagnostic strategies for common medical problems*. Philadelphia: American College of Physicians, 1991.
10 Dinant GJ, Knottnerus JA, Van Wersch JW. Discriminating ability of the erythrocyte sedimentation rate: a prospective study in general practice. *Br J Gen Pract* 1991; **41**:365–70.
11 Stoffers HEJH, Kester ADM, Kaiser V, Rinkens PELM, Knottnerus JA. Diagnostic value of signs and symptoms associated with peripheral arterial obstructive disease seen in general practice: a multivariable approach. *Med Decision Making* 1997;**17**:61–70.
12 Fijten GHF. *Rectal bleeding, a danger signal?* Amsterdam: Thesis Publishers, 1993.
13 Knottnerus JA, Winkens R. Screening and diagnostic tests. In: Silagy C, Haines A, eds. *Evidence based practice in primary care*. London: BMJ Books, 1998.
14 Dinant GJ. *Diagnostic value of the erythrocyte sedimentation rate in general practice*. PhD thesis, University of Maastricht, 1991.
15 Hobbs FD, Delaney BC, Fitzmaurice DA, *et al*. A review of near patient testing in primary care. *Health Technol Assess* 1997;**1**:i–iv,1–229.
16 Campens D, Buntinx F. Selecting the best renal function tests. A meta-analysis of diagnostic studies. *Int J Technol Assess Health Care* 1997;**13**:343–56.
17 van Weel C, Knottnerus JA. Evidence-based interventions and comprehensive treatment. *Lancet* 1999;**353**:916–18.
18 Spiegelhalter DJ, Crean GP, Holden R, *et al*. Taking a calculated risk: predictive scoring systems in dyspepsia. *Scand J Gastroenterol* 1987;**22**(suppl 128):152–60.
19 Knottnerus JA. Diagnostic prediction rules: principles, requirements, and pitfalls. *Primary Care* 1995;**22**:341–63.
20 Knottnerus JA. Application of logistic regression to the analysis of diagnostic data. *Med Decision Making* 1992;**12**:93–108.
21 Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. *BMJ* 1997;**315**:1109–1110.
22 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.
23 Begg CB. Biases in the assessment of diagnostic tests. *Statistics Med* 1987;**6**:411–23.
24 Green MS. The effect of validation group bias on screening tests for coronary artery disease. *Statistics Med* 1985;**4**:53–61.
25 Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;**45**:1143–54.

26 Zarin OA, Pauker SG. Decision analysis as a basis for medical decision making: the tree of Hippokrates. *J Med Philos* 1984;**9**:181–213.
27 Schwartz WB, Wolfe HJ, Pauker SG. Pathology and probabilities, a new approach to interpreting and reporting biopsies. *N Engl J Med* 1981;**305**:917–23.
28 Liechtenstein JI, Feinstein AR, Suzio KD, DeLuca V, Spiro HM. The effectiveness of pandendoscopy on diagnostic and therapeutic decisions about chronic abdominal pain. *J Clin Gastroenterol* 1980;**2**:31–6.
29 Dronfield MW, Langman MJ, Atkinson M, *et al*. Outcome of endoscopy and barium radiography for acute upper gastrointestinal bleeding: controlled trial in 1037 patients. *BMJ* 1982;**284**:545–8.
30 Brett GZ. The value of lung cancer detection by six-monthly chest radiographs. *Thorax* 1968;**23**:414–20.
31 Brown VA, Sawers RS, Parsons RJ, Duncan SL, Cooke ID. The value of antenatal cardiotocography in the management of high risk pregnancy: a randomised controlled trial. *Br J Obster Gynaecol* 1982;**89**:716–22.
32 Flynn AM, Kelly J, Mansfield H, Needham P, O'Connor M, Viegas O. A randomised controlled trial of non-stress antepartum cardiotocography. *Br J Obster Gynaecol* 1982;**89**:427–33.
33 Durbridge TC, Edwards F, Edwards RG, Atkinson M. An evaluation of multiphasic screening on admission to hospital. *Med J Aust* 1976;**1**:703–5.
34 Hull RD, Hirsch J, Carter CJ, *et al*. Diagnostic efficacy of impedance phletysmography for clinically suspected deep-vein thrombosis. A randomized trial. *Ann Intern Med* 1985; **102**:21–8.
35 Shapiro S, Venet W, Strax Ph, Roeser R. Ten to fourteen year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982;**69**:349–55.
36 Tabár L, Fagerberg CJG, Gad A, Badertorp L. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet* 1985;**1**:829–31.
37 Harms LM, Schellevis FG, van Eijk JT, Donker AJ, Bouter LM. Cardiovascular morbidity and mortality among hypertensive patients in general practice: the evaluation of long-term systematic management. *J Clin Epidemiol* 1997;**50**:779–86.
38 Collette HJA, Day NE, Rombach JJ, de Waard F. Evaluation of screening for breast cancer in a non-randomised study (the DOM project) by means of a case control study. *Lancet* 1984;**1**:1224–6.
39 Verbeek ALM, Hendriks JHCL, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of breast cancer mortality through mass-screening with modern mammography. *Lancet* 1984;**1**:1222–4.
40 Guyatt GH, Tugwell P, Feeny DH, Drummond MF, Haynes RB. The role of before–after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chronic Dis* 1986;**39**:295–304.
41 Fineberg HV, Bauman R. Computerized cranial tomography: effect on diagnostic and therapeutic planns. *JAMA* 1977;**238**:224–7.
42 Dinant GJ, Knottnerus JA, van Wersch JW. Diagnostic impact of the erythrocyte sedimentation rate in general practice: a before–after analysis. *Fam Pract* 1991;**9**:28–31.
43 Knottnerus JA. Medical decision making by general practitioners and specialists. *Fam Pract* 1991;**8**:305–7.
44 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–130.
45 Buntinx F, Brouwers M. Relation between sampling device and detection of abnormality in cervical smears: a meta-analysis of randomised and quasi-randomised studies. *BMJ* 1996;**313**:1285–90.
46 Benavente O, Moher D, Pham B. Carotid endarterectomy for asymptomatic carotid stenosis: a meta-analysis. *BMJ* 1998;**317**:1477–80.
47 Van Binsbergen JJ, Brouwer A, van Drenth BB, Haverkort AFM, Prins A, van der Weijden T. Dutch College of General Practitioners standard on cholesterol (M20). *Huisarts Wet* 1991;**34**:551–7.
48 Consensus on non-invasive diagnosis of peripheral arterial vascular disease. Utrecht: CBO, 1994.
49 Feinstein AR. *Clinimetrics*. New Haven: Yale University Press, 1987.

2 The architecture of diagnostic research

DAVID L SACKETT, R BRIAN HAYNES

Summary box

- Because diagnostic testing aims to discriminate between clinically "normal" and "abnormal", the definition of "normal" and "the normal range" is a basic issue in diagnostic research. Although the "gaussian" definition is traditionally common, the "therapeutic definition" of normal is the most clinically relevant.
- The diagnostic research question to be answered has to be carefully formulated, and determines the appropriate research approach. The four most relevant types of question are:
- **Phase I questions: Do patients with the target disorder have different test results from normal individuals?** The answer requires a comparison of the distribution of test results among patients known to have the disease and people known not to have the disease.
- **Phase II questions: Are patients with certain test results more likely to have the target disorder than patients with other test results?** This can be studied in the same dataset that generated the Phase I answer, but now test characteristics such as sensitivity and specificity are estimated.
- Only if Phase I and Phase II studies, performed in "ideal circumstances", are sufficiently promising as to possible discrimination between diseased and non-diseased subjects, it is worth evaluating the test under "usual" circumstances. Phase III and IV questions must then be answered.
- **Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the test result**

distinguish those with and without the target disorder? To get the appropriate answer, a consecutive series of such patients should be studied.
- The validity of Phase III studies is threatened when cases where the reference standard or diagnostic test is lost, not performed, or indeterminate, are frequent or inappropriately dealt with.
- Because of a varying patient mix, test characteristics such as sensitivity, specificity and likelihood ratios may vary between different healthcare settings.
- **Phase IV questions: Do patients who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not?** These questions have to be answered by randomising patients to undergo the test of interest or some other (or no) test.

Introduction

When making a diagnosis, clinicians seldom have access to reference or "gold" standard tests for the target disorders they suspect, and often wish to avoid the risks or costs of these reference standards, especially when they are invasive, painful, or dangerous. No wonder, then, that clinical researchers examine relationships between a wide range of more easily measured phenomena and final diagnoses. These phenomena include elements of the patient's history, physical examination, images from all sorts of penetrating waves, and the levels of myriad constituents of body fluids and tissues. Alas, even the most promising phenomena, when nominated as diagnostic tests, almost never exhibit a one-to-one relationship with their respective target disorders, and several different diagnostic tests may compete for primacy in diagnosing the same target disorder. As a result, considerable effort has been expended at the interface between clinical medicine and scientific methods in an effort to maximise the validity and usefulness of diagnostic tests. This book describes the result of those efforts, and this chapter focuses on the specific sorts of questions posed in diagnostic research and the study architectures used to answer them.

At the time that this book was being written, considerable interest was being directed to questions about the usefulness of the plasma concentration of B-type natriuretic peptide in diagnosing left ventricular dysfunction.[1] These questions were justified on two grounds: first, left ventricular dysfunction is difficult to diagnose on clinical examination; and second, randomised trials have shown that treating it (with angiotensin

converting enzyme inhibitors) reduces its morbidity and mortality. Because real examples are far better than hypothetical ones in illustrating not just the overall strategies but also the down-to-earth tactics of clinical research, we will employ this one in the following paragraphs. To save space and tongue twisting we will refer to the diagnostic test, B-type natriuretic peptide, as BNP and the target disorder it is intended to diagnose, left ventricular dysfunction, as LVD. The starting point in evaluating this or any other promising diagnostic test is to decide how we will define its normal range.

What do you mean by "normal" and "the normal range"?

This chapter deals with the strategies (a lot) and tactics (a little) of research that attempts to distinguish patients who are "normal" from those who have a specific target disorder. Before we begin, however, we need to acknowledge that several different definitions of normal are used in clinical medicine, and we confuse them at our (and patients') peril. We know six of them[2] and credit Tony Murphy for pointing out five.[3] A common "*gaussian*" definition (fortunately falling into disuse) assumes that the diagnostic test results for BNP (or some arithmetic manipulation of them) for everyone, or for a group of presumably normal people, or for a carefully characterised "reference" population, will fit a specific theoretical distribution known as the *normal* or *gaussian* distribution. Because the mean of a gaussian distribution plus or minus 2 standard deviations encloses 95% of its contents, it became a tempting way to define the normal several years ago, and came into general use. It is unfortunate that it did, for three logical consequences of its use have led to enormous confusion and the creation of a new field of medicine: the diagnosis of non-disease. First, diagnostic test results simply do not fit the gaussian distribution (actually, we should be grateful that they do not; the gaussian distribution extends to infinity in both directions, necessitating occasional patients with impossibly high BNP results and others on the minus side of zero!). Second, if the highest and lowest 2.5% of diagnostic test results are called abnormal, then all the diseases they represent have exactly the same frequency, a conclusion that is also clinically nonsensical.

The third harmful consequence of the use of the gaussian definition of normal is shared by its more recent replacement, the *percentile*. Recognising the failure of diagnostic test results to fit a theoretical distribution such as the gaussian, some laboratorians have suggested that we ignore the shape of the distribution and simply refer (for example) to the lower (or upper) 95% of BNP or other test results as normal. Although this percentile definition does avoid the problems of infinite and negative test values, it still suggests

that the underlying prevalence of all diseases is similar – about 5% – which is silly, and still contributes to the "upper-limit syndrome" of non-disease because its use means that the only "normal" patients are the ones who are not yet sufficiently worked up. This inevitable consequence arises as follows: if the normal range for a given diagnostic test is defined as including the lower 95% of its results, then the probability that a given patient will be called "normal" when subjected to this test is 95%, or 0.95. If this same patient undergoes two independent diagnostic tests (independent in the sense that they are probing totally different organs or functions), the likelihood of this patient being called normal is now $(0.95) \times (0.95) = 0.90$. So, the likelihood of any patient being called normal is 0.95 raised to the power of the number of independent diagnostic tests performed on them. Thus, a patient who undergoes 20 tests has only 0.95 to the 20th power, or about one chance in three, of being called normal; a patient undergoing 100 such tests has only about six chances in 1000 of being called normal at the end of the work up.*

Other definitions of normal, in avoiding the foregoing pitfalls, present other problems. The *risk factor* definition is based on studies of precursors or statistical predictors of subsequent clinical events; by this definition, the normal range for BNP or serum cholesterol or blood pressure consists of those levels that carry no additional risk of morbidity or mortality. Unfortunately, however, many of these risk factors exhibit steady increases in risk throughout their range of values; indeed, some hold that the "normal" total serum cholesterol (defined by cardiovascular risk) might lie well below 3.9 mmol/L (150 mg%), whereas our local laboratories employ an upper limit of normal of 5.2 mmol/L (200 mg%), and other institutions employ still other definitions.

Another shortcoming of this risk factor definition becomes apparent when we examine the health consequences of acting upon a test result that lies beyond the normal range: will altering BNP or any other risk factor really change risk? For example, although obesity is a risk factor for hypertension, controversy continues over whether weight reduction improves mild hypertension. One of us led a randomised trial in which we peeled 4.1 kg (on average) from obese, mildly hypertensive women with a behaviourally oriented weight reduction programme the (control women lost less than 1 kg).[4] Despite both their and our efforts (the cost of the experimental group's behaviourally oriented weight reduction programme came to US$60 per kilo), there was no accompanying decline in blood pressure.

A related approach defines the normal as that which is *culturally desirable*, providing an opportunity for what HL Mencken called "the corruption of

*This consequence of such definitions helps explain the results of a randomised trial of hospital admission multitest screening that found no patient benefits, but increased healthcare costs, when such screening was carried out.[20]

medicine by morality" through the "confusion of the theory of the healthy with the theory of the virtuous".[5] Although this definition does not fit our BNP example, one sees such definitions in their mostly benign form at the fringes of the current lifestyle movement (for example, "It is better to be slim than fat,"† and "Exercise and fitness are better than sedentary living and lack of fitness"), and in its malignant form in the healthcare system of the Third Reich. Such a definition has the potential for considerable harm, and may also serve to subvert the role of medicine in society.

Two final definitions are highly relevant and useful to the clinician because they focus directly on the clinical acts of diagnosis and therapy. The *diagnostic* definition identifies a range of BNP (or other diagnostic test) results beyond which LVD (or another specific target disorder) is (with known probability) present. It is this definition that we focus on in this book. The "known probability" with which a target disorder is present is known formally as the positive predictive value, and depends on where we set the limits for the normal range of diagnostic test results. This definition has real clinical value and is a distinct improvement over the definitions described above. It does, however, require that clinicians keep track of diagnostic ranges and cut-offs.

The final definition of normal sets its limits at the level of BNP beyond which specific treatments for LVD (such as ACE inhibitors) have been shown conclusively to do more good than harm. This *therapeutic* definition is attractive because of its link with action. The therapeutic definition of the normal range of blood pressure, for example, avoids the hazards of labelling patients as diseased unless they are going to be treated. Thus, in the early 1960s the only levels of blood pressure conclusively shown to benefit from antihypertensive drugs were diastolic pressures in excess of 130 mmHg (phase V). Then, in 1967, the first of a series of randomised trials demonstrated the clear advantages of initiating drugs at 115 mmHg, and the upper limit of normal blood pressure, under the therapeutic definition, fell to that level. In 1970 it was lowered further to 105 mmHg with a second convincing trial, and current guidelines about which patients have abnormal blood pressures that require treatment add an element of the risk factor definition and recommend treatment based on the combination of blood pressure with age, sex, cholesterol level, blood sugar, and smoking habit. These days one can even obtain evidence for blood pressure treatment levels based on the presence of a second disease: for example, in type 2 diabetes the "tight control" of blood pressure reduces the risk of major complications in a cost effective way. Obviously, the use of this therapeutic definition requires that clinicians (and guideline developers) keep abreast of advances in therapeutics, and that is as it should be.

†But the tragic consequences of anorexia nervosa teach us that even this definition can do harm.

In summary, then, before you start any diagnostic study you need to define what you mean by normal, and be confident that you have done so in a sensible and clinically useful fashion.

The question is everything

As in other forms of clinical research, there are several different ways in which one could carry out a study into the potential or real diagnostic usefulness of a physical sign or laboratory test, and each of them is appropriate to one sort of question and inappropriate for others. Among the questions one might pose about the relation between a putative diagnostic test (say, BNP) and a target disorder (say, LVD), four are most relevant:

- **Phase I questions**: Do patients with the target disorder have different test results from normal individuals? (Do patients with LVD have higher BNP than normal individuals?)
- **Phase II questions**: Are patients with certain test results more likely to have the target disorder than patients with other test results? (Are patients with higher BNP more likely to have LVD than patients with lower BNP?)
- **Phase III questions**: Among patients in whom it is clinically sensible to suspect the target disorder, does the level of the test result distinguish those with and without the target disorder? (Among patients in whom it is clinically sensible to suspect LVD, does the level of BNP distinguish those with and without LVD?)
- **Phase IV questions**: Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not? (Of greatest interest in evaluating early diagnosis through screening tests, this might be phrased: Do patients screened with BNP (in the hope of achieving the early diagnosis of LVD) have better health outcomes (mortality, function, quality of life) than those who do not undergo screening?).

At first glance the first three questions may appear indistinguishable or even identical. They are not, because the strategies and tactics employed in answering them are crucially different, and so are the conclusions that can be drawn from their answers. The first two differ in the "direction" in which their results are analysed and interpreted, and the third differs from the first two as well in the fashion in which study patients are assembled. The fourth question gets at what we and our patients would most like to know: are they better off for having undergone it? The conclusions that can (and, more importantly, cannot) be drawn from the answers to these questions are crucially different, and there are plenty of examples of the price paid by patients and providers when the answers to Phase I or II questions are

interpreted as if they were answering a Phase III (or even a Phase IV) question.

These questions also nicely describe an orderly and efficient progression of research into the potential usefulness of a clinical sign, symptom, or laboratory result, and we will use the BNP story to show this sequence.

Phase I questions: Do patients with the target disorder have different test results from normal individuals?

Question 1 often can be answered with a minimum of effort, time, and expense, and its architecture is displayed in Table 2.1.

For example, a group of investigators at a British university hospital measured BNP precursor in convenience samples of "normal controls" and in patients who had various combinations of hypertension, ventricular hypertrophy, and LVD.[6] They found statistically significant differences in median BNP precursors between patients with and normal individuals without LVD, and no overlap in their range of BNP precursor results. It was not surprising, therefore, that they concluded that BNP was "a useful diagnostic aid for LVD".

Note, however, that the direction of interpretation here is from known diagnosis back to diagnostic test. Answers to Phase I questions cannot be applied directly to patients because they are presented as overall (usually average) test results. They are not analysed in terms of the diagnostic test's sensitivity, specificity, or likelihood ratios. Moreover, Phase I studies are typically conducted among patients known to have the disease and people known not to have the disease (rather than among patients who are suspected of having, but not known to have, the disease). As a result, this phase of diagnostic test evaluation cannot be translated into diagnostic action.

Why, then, ask Phase I questions at all? There are two reasons. First, such studies add to our biologic insights about the mechanisms of disease, and may serve later research into therapy as well as diagnosis. Second, such studies are quick and relatively cheap, and a negative answer to their question removes the need to ask the tougher, more time-consuming, and costlier questions of Phases II–IV. Thus, if a convenience (or "grab")

Table 2.1 Answering a Phase I question: Do patients with LVD have higher BNP than normal individuals?

	Patients known to have the target disorder (LVD)	Normal controls
Average diagnostic test (BNP precursor) result (and its range)	493.5 (range from 248.9 to 909)	129.4 (range from 53.6 to 159.7)

sample of patients with LVD already known to the investigators displays the same average levels and distribution of BNP as apparently healthy laboratory technicians or captive medical students, it is time to abandon it as a diagnostic test and devote scarce resources to some other lead.

Phase II questions: Are patients with certain test results more likely to have the target disorder than patients with other test results?

Following a positive answer to a Phase I question, it is logical to ask a Phase II question, this time changing the direction of interpretation so that it runs from diagnostic test result forward to diagnosis. Although the Phase II questions often can be asked in the same dataset that generated the Phase I answer, the architecture of asking and answering them differs. For example, a second group of investigators at a Belgian university hospital measured BNP in "normal subjects" and 3 groups of patients with coronary artery disease and varying degrees of LVD.[7] Among the analyses they performed (including the creation of ROC curves; see Chapter 7) was a simple plot of individual BNP results, generating the results shown in Table 2.2 by picking the cut-off that best distinguished their patients with severe LVD from their normal controls.

As you can see, the results in Table 2.2 are extremely encouraging. Whether it is used to "rule out" LVD on the basis of its high sensitivity (SnNout)[8] or to "rule in" LVD with its high specificity (SpPin),[9] BNP looks useful, so it is no wonder that the authors concluded: "BNP concentrations are good indicators of the severity and prognosis of

Table 2.2 Answering a Phase II question: Are patients with higher BNP more likely to have LVD than patients with lower BNP?

	Patients known to have the target disorder (LVD)	Normal controls
High BNP	39	2
Normal BNP	1	25
Test characteristics and their 95% confidence intervals	Lower	Upper
Sensitivity = 98%	87%	100%
Specificity = 92%	77%	98%
Positive predictive value = 95%	84%	99%
Negative predictive value = 96%	81%	100%
Likelihood ratio for an abnormal test result = 13	3.5	50
Likelihood ratio for a normal test result = 0.03	0.0003	0.19

congestive heart failure". But is Table 2.2 overly encouraging? It compares test results between groups of patients who already have established diagnoses (rather than those who are merely suspected of the target disorder), and contrasts extreme groups of normals and those with severe disease. Thus, it tells us whether the test shows diagnostic promise under ideal conditions. A useful way to think about this difference between Phase II

Table 2.3 Explanatory and pragmatic studies of diagnostic tests and treatments.

	Promising diagnostic test		Promising treatment	
Feature	Explanatory (Phase II study)	Pragmatic (Phase III study)	Explanatory	Pragmatic
Question	Can this test discriminate under ideal circumstances?	Does this test discriminate in routine practice?	Efficacy: Can this treatment work under ideal circumstances?	Effectiveness: Does this treatment work in routine practice?
Selection of patients	Preselected groups of normal individuals and of those who clearly have the target disorder	Consecutive patients in whom it is clinically sensible to suspect the target disorder	Highly compliant, high-risk, high-response patients	All comers, regardless of compliance, risk or responsiveness
Application of manoeuvre	Carried out by expert clinician or operator on best equipment	Carried out by usual clinician or operator on usual equipment	Administered by experts with great attention to compliance	Administered by usual clinicians under usual circumstances
Definition of outcomes	Same reference standard for those with and without the target disorder	Often different standards for patients with and without the target disorder; may invoke good treatment-free prognosis as proof of absence of target disorder	May focus on pathophysiology, surrogate outcomes, or cause-specific mortality	"Hard" clinical events or death (often all-cause mortality)
Exclusion of patients or events	Often exclude patients with lost results and indeterminate diagnoses	Include all patients, regardless of lost results or indeterminate diagnoses	May exclude events before or after treatment is applied	Includes all events after randomisation
Results confirmed in a second, independent ("test") sample of patients	Usually not	Ideally yes		
Incorporation into systematic review	Usually not	Ideally yes	Sometimes	Ideal

and Phase III studies is by analogy with randomised clinical trials, which range from addressing explanatory (efficacy) issues of therapy (can the new treatment work under ideal circumstances?) to management (pragmatic, effectiveness) issues (does the new treatment work under usual circumstances?). We have summarised this analogy in Table 2.3.

As shown in Table 2.3, the Phase II study summarised in Table 2.2 is explanatory in nature: preselected groups of normal individuals (ducks) and those who clearly have the target disorder (yaks) undergo testing under the most rigorous circumstances possible, with the presence or absence of the target disorder being determined by the same reference standard. No attempt is made to validate these initial ("training set") results (especially the cut-off used to set the upper limit of normal BNP) in a second, independent "test" set of ducks and yaks. On the other hand, and as with the Phase I study, this relatively easy Phase II investigation tells us whether the promising diagnostic test is worth further, costlier evaluation; as we have said elsewhere,[10] if the test cannot tell the difference between a duck and a yak it is worthless in diagnosing either one. As long as the writers and readers of a Phase II explanatory study report make no pragmatic claims about its usefulness in routine clinical practice, no harm is done. Furthermore, criticisms of Phase II explanatory studies for their failure to satisfy the methodological standards employed in Phase III pragmatic studies do not make sense.

Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the level of the test result distinguish those with and without the target disorder?

Given its promise in Phase I and II studies, it is understandable that BNP would be tested in the much costlier and more time-consuming Phase III study, in order to determine whether it was really useful among patients in whom it is clinically sensible to suspect LVD. As we were writing this chapter, an Oxfordshire group of clinical investigators reported that they did just that by inviting area general practitioners "to refer patients with suspected heart failure to our clinic".[11] Once there, these 126 patients underwent independent, blind BNP measurements and echocardiography. Their results are summarised in Table 2.4.

About one third of the patients referred by their general practitioners had LVD on echocardiography. These investigators documented that BNP measurements did not look nearly as promising when tested in a Phase III study in the pragmatic real-world setting of routine clinical practice, and concluded that "introducing routine measurement [of BNP] would be unlikely to improve the diagnosis of symptomatic [LVD] in the

Table 2.4 Answering a Phase III question: Among patients in whom it is clinically sensible to suspect LVD, does the level of BNP distinguish patients with and without LVD?

	Patients with LVD on echocardiography	Patients with normal echoes
High BNP (>17.9 pg/ml)	35	57
Normal BNP (<18 pg/ml)	5	29
Prevalence or pretest probability of LVD	40/126 = 32%	
Test characteristics and their 95% confidence intervals	Lower	Upper
Sensitivity = 88%	74%	94%
Specificity = 34%	25%	44%
Positive predictive value = 38%	29%	48%
Negative predictive value = 85%	70%	94%
Likelihood ratio for an abnormal test result = 1.3	1.1	1.6
Likelihood ratio for a normal test result = 0.4	0.2	0.9

Table 2.5 Answering a Phase III question with likelihood ratios.

	Patients with LVD on echocardiography	Patients with normal echoes	Likelihood ratio and 95% CI
High BNP (>76 pg/ml)	26 (0.650)	11 (0.128)	5.1 (2.8–9.2)
Mid BNP (10–75 pg/ml)	11 (0.275)	60 (0.698)	0.4 (0.2–0.7)
Low BNP (<10 pg/ml)	3 (0.075)	15 (0.174)	0.4 (0.1–1)
Total	40 (1.000)	86 (1.000)	

community". However, their report of the study also documented the effect of two other cut-points for BNP. This led both to a counterclaim on the usefulness of BNP in the subsequent email letters to the editor, and to an opportunity for us to describe an alternative way of presenting information about the accuracy of a diagnostic test: the multilevel likelihood ratio (LR). The original report makes it possible for us to construct Table 2.5.

By using multilevel likelihood ratios to take advantage of the full range of BNP results, we can be slightly more optimistic about the diagnostic usefulness of higher levels: the LR for BNP results >76 pg/ml was 5.1. These levels were found in 29% of the patients in this study, and their presence raised the pretest probability of LVD in the average patient from 32% to a post-test probability of 70%. This can be determined directly from Table 2.5 for this "average" patient with a pretest probability of 32% and a high BNP: reading horizontally across the top row, the result is 26/(26+11) = 70%.

However, if the patient has a different pretest likelihood, say 50%, then either the table must be reconstructed for this higher figure, or the pretest probability needs to be converted to a pretest odds (50% is a pretest odds of $(1-0.5)/0.5 = 1$), and then multiplied by the likelihood ratio for the test result (5.1 in this case), giving a post-test odds of 5.1, which then can be converted back into a post-test probability of $5.1/(1+5.1) = 84\%$. These calculations are rendered unnecessary by using a nomogram, as in Figure 2.1.

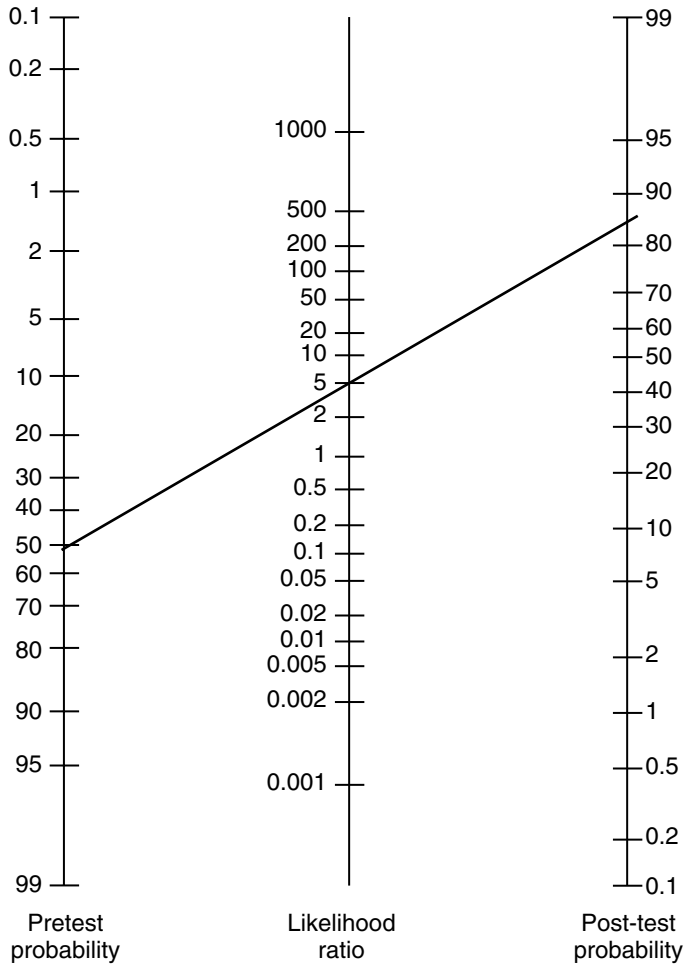


Figure 2.1 Nomogram for converting pretest likelihoods (left column) to post-test likelihoods (right column) by drawing a straight line from the pretest likelihood through the likelihood ratio for the test result.

Given the quite wide confidence intervals around these LRs, further type III studies may be fruitful (and readers can check to see whether this was done after this chapter was published).

Threats to the validity of Phase III studies

There are several threats to the validity of Phase III studies that distort their estimates of the accuracy of the diagnostic test, and the first batch are violations of the old critical appraisal guide: "Has there been an independent, blind comparison with a gold standard of diagnosis?"[12] By *independence* we mean that *all* study patients have undergone *both* the diagnostic test *and* the reference ("gold") standard evaluation and, more specifically, that the reference standard is applied *regardless of the diagnostic test result*. By *blind* we mean that the reference standard is applied and interpreted in total ignorance of the diagnostic test result, and vice versa. By anticipating these threats at the initial question forming phase of a study, they can be avoided or minimised.

Although we prefer to conceptualise diagnostic test evaluations in terms of 2×2 tables such as the upper panel of Table 2.6 (and this is the way that most Phase II studies are performed), in reality Phase III studies generate the 3×3 tables shown in the lower panel of Table 2.6. Reports get lost, their results are sometimes incapable of interpretation, and sometimes we are unwilling to apply the reference standard to all the study patients.

The magnitude of the cells ***v***–***z*** and the method of handling patients who fall into these cells will affect the validity of the study. In the perfect study these cells are kept empty, or so small that they cannot exert any important

Table 2.6 The ideal Phase III study meets the real world.

	Reference standard	
The ideal study	Target disorder present	Target disorder absent
Diagnostic test result		
Positive	***a***	***b***
Negative	***c***	***d***

	Reference standard		
The real study	Target disorder present	Lost, not performed, or indeterminate	Target disorder absent
Diagnostic test result			
Positive	***a***	***v***	***b***
Lost, not performed, or indeterminate	***w***	***x***	***y***
Negative	***c***	***z***	***d***

effect on the study conclusions. However, there are 6 situations in which they become large enough to bias the measures of test accuracy. First, when the reference standard is expensive, painful, or risky, investigators will not wish to apply it to patients with negative diagnostic test results. As a consequence, such patients risk winding up in cell ***z***. Furthermore, there is an understandable temptation to shift them to cell ***d*** in the analysis. Because no diagnostic test is perfect, some of them surely belong in cell ***c***. Shifting all of them to cell ***d*** falsely inflates both sensitivity and specificity. If this potential problem is recognised before the study begins, investigators can design their reference standard to prevent such patients from falling into cell ***z***. This is accomplished by moving to a more pragmatic study and adding another, prognostic dimension to the reference standard, namely the clinical course of patients with negative test results who receive no intervention for the target disorder. If patients who otherwise would end up in cell ***z*** develop the target disorder during this treatment-free follow up, they belong in cell ***c***. If they remain free of disease, they join cell ***d***. The result is an unbiased and pragmatic estimate of sensitivity and specificity.

Second, the reference standard may be lost; and third, it may generate an uninterpretable or indeterminate result. As before, arbitrarily analysing such patients as if they really did or did not have the target disorder will distort measures of diagnostic test accuracy. Once again, if these potential biases are identified in the planning stages they can be minimised, a pragmatic solution such as that proposed above for cell ***z*** considered, and clinically sensible rules established for shifting them to the definitive columns in a manner that confers the greatest benefit (in terms of treatment) and the least harm (in terms of labelling) to later patients.

Fourth, fifth, and sixth, the diagnostic test result may be lost, never performed, or indeterminate, so that the patient winds up in cells ***w***, ***x***, or ***y***. Here the only unforgivable action is to exclude such patients from the analysis of accuracy. As before, anticipation of these problems before the study begins should minimise tests that are lost or never performed to the point where they would not affect the study conclusion regardless of how they were classified. If indeterminate results are likely to be frequent, a decision can be made before the study begins as to whether they will be classified as positive or negative. Alternatively, if multilevel likelihood ratios are to be used, these patients can form their own stratum.

In addition to the 6 threats to validity related to cells ***v–z***, there are two more. The seventh threat to validity noted in the above critical appraisal guide arises when a patient's reference standard is applied or interpreted by someone who already knows that patient's diagnostic test result (and vice versa). This is a risk whenever there is any degree of interpretation (even in reading off a scale) involved in generating the result of the diagnostic test or reference standard. We know that these situations lead to biased inflations of sensitivity and specificity.

The eighth and final threat to the validity of accuracy estimates generated in Phase III studies arises whenever the selection of the "upper limit of normal" or cut-point for the diagnostic test is under the control of the investigator. When they can place the cut-point wherever they want, it is natural for them to select the point where it maximises sensitivity (for use as a SnNout), specificity (for use as a SpPin), or the total number of patients correctly classified in that particular "training" set. If the study were repeated in a second, independent "test" set of patients, employing that same cut-point, the diagnostic test would be found to function a little or a lot worse. Thus, the true accuracy of a promising diagnostic test is not known until it has been evaluated in one or more independent studies.

The foregoing threats apply whether the diagnostic test comprises a single measurement of a single phenomenon or a multivariate combination of several phenomena. For example, Philip Wells and his colleagues determined the diagnostic accuracy of the combination of several items from the medical history, physical examination, and non-invasive testing in the diagnosis of deep vein thrombosis.[13] Although their study generated similar results in three different centres (two in Canada and one in Italy), even they recommended further prospective testing before widespread use.

Limits to the applicability of Phase III studies

Introductory courses in epidemiology introduce the concept that predictive values change as we move back and forth between screening or primary care settings (with their low prevalence or pretest probability of the target disorder) to secondary and tertiary care (with their higher probability of the target disorder). This point is usually made by assuming that sensitivity and specificity remain constant across all settings. However, the mix (or spectrum) of patients also varies between these locations; for example, screening is applied to asymptomatic individuals with early disease, whereas tertiary care settings deal with patients with advanced or florid disease. No wonder, then, that sensitivity and specificity often vary between these settings. Moreover, because primary care patients with positive diagnostic test results (which comprise false positive as well as true positive results) are referred forward to secondary and tertiary care, we might expect specificity to fall as we move along the referral pathway. There is very little empirical evidence addressing this issue, and we acknowledge our debt to Dr James Wagner of the University of Texas at Dallas for tracking down and systematically reviewing diagnostic data from over 2000 patients with clinically suspected appendicitis seen in primary care and on inpatient surgical wards (personal communication, 2000). The diagnostic tests comprised the clinical signs that are sought when clinicians suspect appendicitis, and the reference standard is a combination of pathology

Table 2.7 The accuracy of right lower quadrant tenderness in the diagnosis of appendicitis.

	Primary care settings Appendicitis		Tertiary care settings Appendicitis	
	Yes (%)	No (%)	Yes (%)	No (%)
Right lower quadrant tenderness				
Present	84	11	81	84
Absent	16	89	19	16
Total	100	100	100	100
Frequency of appendicitis	14%		63%	
Frequency of positive sign	21%		82%	
Sensitivity	84%		81%	
Specificity	89%		16%	
LR+	7.6		1	
LR−	0.2		1	

reports on appendices when operations were performed, and a benign clinical course when they were not. The results for the diagnostic test of right lower quadrant tenderness are shown in Table 2.7.

A comparison of the results in primary and tertiary care shows, as we might expect, an increase in the proportions of patients with appendicitis (from 14% to 63%). But, of course, this increase in prevalence occurred partly because patients with right lower quadrant tenderness (regardless of whether this was a true positive or false positive finding) tended to be referred to the next level of care, whereas patients without this sign tended not to be referred onward; this is confirmed by the rise in the frequency of this sign from 21% of patients in primary care to 82% of patients in tertiary care. Although this sort of increase in a positive diagnostic test result is widely recognised, its effect on the accuracy of the test is not. The forward referral of patients with false positive test results leads to a fall in specificity, in this case a dramatic one from 89% down to 16%. As a result, a diagnostic sign of real value in primary care (LR+ of 8, LR− of 0.2) is useless in tertiary care (LR+ and LR− both 1); in other words, its diagnostic value has been "used up" along the way.‡

This phenomenon can place major limitations on the applicability of Phase III studies carried out in one sort of setting to another setting where

‡Although not germane to this book on research methods, there are two major clinical ramifications of this phenomenon. First, because clinical signs and other diagnostic tests often lose their value along the referral pathway, tertiary care clinicians might be forgiven for proceeding immediately to applying invasive reference standards. Second, tertiary care teachers should be careful what they teach primary care trainees about the uselessness of clinical signs.

Table 2.8 The accuracy of abdominal rigidity in the diagnosis of appendicitis.

	Primary care settings Appendicitis		Tertiary care settings Appendicitis	
	Yes (%)	No (%)	Yes (%)	No (%)
Rigid abdomen				
Present	40	26	23	6
Absent	60	74	77	94
Total	100	100	100	100
Frequency of appendicitis	14%		47%	
Frequency of positive sign	28%		14%	
Sensitivity	40%		24%	
Specificity	74%		94%	
LR+	1.5		5	
LR−	0.8		0.8	

the mix of test results may differ. Overcoming this limitation is another bonus that attends the replication of a promising Phase III study in a second "test" setting attended by patients of the sort that the test is claimed to benefit.

Does specificity always fall between primary care and tertiary care settings? Might this be employed to generate a "standardised correction factor" for extrapolating test accuracy between settings? Have a look at the clinical sign of abdominal rigidity in Table 2.8.

In this case, a clinical sign that is useless in primary care (LR+ barely above 1 and LR− close to 1) is highly useful in tertiary care (LR+ of 5), and in this case specificity has risen (from 74% to 95%), not fallen, along the referral pathway. The solution to this paradox is revealed in the frequency of the sign in these two settings; it has fallen (from 28% to 14%), not risen, along the pathway from primary to tertiary care. We think that the explanation is that primary care clinicians, who do not want to miss any patient's appendicitis, are "over-reading" abdominal rigidity compared to their colleagues in tertiary care. At this stage in our knowledge of this phenomenon we do not think the "standard correction factors" noted in the previous paragraph are advisable, and this paradox once again points to the need to replicate promising Phase III study results in "test" settings attended by patients (and clinicians!) of the sort that the test is claimed to benefit. In this regard we welcome the creation of the CARE consortium of over 800 clinicians from over 70 countries[14] for their performance of web-based, large, simple, fast studies of the clinical examination.[15] It is hoped that this group, which can be contacted at www.carestudy.com, can make a large contribution to determining the wide applicability of the

diagnostic test information obtained from the medical history and physical examination.

For clinicians who wish to apply the bayesian properties of diagnostic tests, accurate estimates of the pretest probability of target disorders in their locale and setting are required. These can come from five sources: personal experience, population prevalence statistics, practice databases, the publication that described the test, or one of a growing number of primary studies of pretest probability in different settings.[16]

Phase IV questions: Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not?

The ultimate value of a diagnostic test is measured in the health outcomes produced by the further diagnostic and therapeutic interventions it precipitates. Sometimes this benefit is self-evident, as in the correct diagnosis of patients with life threatening target disorders who thereby receive life saving treatments. At other times these outcomes can be hinted at in Phase III studies if the reference standard for the absence of the target disorder is a benign clinical course despite the withholding of treatment. More often, however, Phase IV questions are posed about diagnostic tests that achieve the early detection of asymptomatic disease, and can only be answered by the follow up of patients randomised to undergo the diagnostic test of interest or some other (or no) test.

Methods for conducting randomised trials are discussed elsewhere,[17] and we will confine this discussion to an example of the most powerful sort, a systematic review of several randomised trials of faecal occult blood testing.[18] In these trials, over 400 000 patients were randomised to undergo annual or biennial screening or no screening, and then carefully followed for up to 13 years in order to determine their mortality from colorectal cancer. The results are summarised in Table 2.9.

Table 2.9 A systematic review of randomised trials of screening for colorectal cancer.

Outcome	Unscreened group	Screened group	Relative risk reduction	Absolute risk reduction	Number needed to screen to prevent one more colorectal cancer death
Colorectal cancer mortality	0.58%	0.50%	16%	0.08%	1237

In this example, patients were randomised to undergo or not undergo the diagnostic test. Because most of them remained cancer free, the sample size requirement was huge and the study architecture is relatively inefficient. It would have been far more efficient (but unacceptable) to randomise the disclosure of positive test results, and this latter strategy was employed in a randomised trial of a developmental screening test in childhood.[19] In this study, the experimental children whose positive test results were revealed and who subsequently received the best available counselling and interventions fared no better in their subsequent academic, cognitive or developmental performance than control children whose positive test results were concealed. However, parents of the "labelled" experimental children were more likely to worry about their school performance, and their teachers tended to report more behavioural problems among them. This warning that diagnostic tests can harm as well as help those who undergo them is a suitable stopping point for this chapter.

References

1 Hobbs R. Can heart failure be diagnosed in primary care? *BMJ* 2000;**321**:188–9.
2 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: a Basic Science for Clinical Medicine*, 2nd edn. Boston: Little, Brown and Company, 1991; pp. 58–61.
3 Murphy EA. The normal, and perils of the sylleptic argument. *Perspect Biol Med* 1972;**15**:566.
4 Haynes RB, Harper AC, Costley SR, *et al*. Failure of weight reduction to reduce mildly elevated blood pressure: a randomized trial. *J Hypertension* 1984;**2**:535.
5 Mencken HL. *A Mencken chrestomathy*. Westminster: Knopf, 1949;12.
6 Talwar S, Siebenhofer A, Williams B, Ng L. Influence of hypertension, left ventricular hypertrophy, and left ventricular systolic dysfunction on plasma N terminal pre-BNP. *Heart* 2000;**83**:278–82.
7 Selvais PL, Donickier JE, Robert A, *et al*. Cardiac natriuretic peptides for diagnosis and risk stratification in heart failure. *Eur J Clin Invest* 1998;**28**:636–42.
8 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 83.
9 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 77.
10 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 57.
11 Landray MJ, Lehman R, Arnold I. Measuring brain natriuretic peptide in suspected left ventricular systolic dysfunction in general practice: cross-sectional study. *BMJ* 2000; **320**:985–6.
12 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 52.
13 Wells PS, Hirsh J, Anderson DR, *et al*. A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process. *J Intern Med* 1998;**243**:15–23.
14 McAlister FA, Straus SE, Sackett DL, on behalf of the CARE-COAD group. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. *Lancet* 1999;**354**:1721–4.
15 Straus SE, McAlister FA, Sackett DL, Deeks JJ. The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease. *JAMA* 2000; **283**:1853–7.
16 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: how to Practise and Teach EBM*, 2nd edn. Edinburgh: Churchill Livingstone, 2000;82–4.
17 Shapiro SH, Louis TA. *Clinical Trials: Issues and Approaches*, 2nd edn. New York: Marcel Dekker, 2000; (in press).

18 Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. *Screening for colorectal cancer using the faecal occult blood test, Hemoccult.* Cochrane Review, latest version 16 Jan 1998. In: The Cochrane Library, Oxford: Update Software.
19 Cadman D, Chambers LW, Walter SD, Ferguson R, Johnston N, McNamee J. Evaluation of public health preschool child development screening: The process and outcomes of a community program. *Am J Public Health* 1987;**77**:45–51.
20 Dunbridge TC, Edwards F, Edwards RG, Atkinson M. *An evaluation of multiphasic screening on admission to hospital.* Précis of a report to the National Health and Medical Research Council. *Med J Aust* 1976;**1**:703–5.

3 Assessment of the accuracy of diagnostic tests: the cross-sectional study

J ANDRÉ KNOTTNERUS, JEAN W MURIS

Summary box

- In determining the accuracy of diagnostic testing, the first step is to appropriately define the contrast to be evaluated. Options are: to evaluate one single test contrast; to compare two or more single tests; to evaluate further testing in addition to previous diagnostics; and to compare alternative diagnostic strategies. In addition, the clinical diagnostic problem under study must be specified. Finally, distinction should be made between evaluating testing in "extreme contrast" or "clinical practice" settings.
- For accuracy studies, general design types are (1) a survey of the total study population, (2) a case–referent approach, or (3) a test-based enrolment. The direction of the data collection should generally be prospective, but ambispective and retrospective approaches are sometimes appropriate.
- One should specify the determinants of primary interest (the test(s) under study) and other potentially important determinants (possible modifiers of test accuracy and confounding variables).
- The reference standard procedure to measure the target disorder should be applied on all included subjects, independently of the result of the test under study. Applying a reference standard procedure can be difficult because of classification errors, lack of a

well defined pathophysiological concept, incorporation bias, or too invasive or too complex patient investigations. Possible solutions are: an independent expert panel, and the delayed-type cross-sectional study (clinical follow up). Also, a prognostic criterion can be chosen to define clinical outcome.

- Inclusion criteria must be based on "the intention to diagnose" or "intention to screen" with respect to the studied clinical problem. The recruitment procedure is preferably a consecutive series of presenting patients or a target population screening, respectively.
- In the design phase, sample size estimation should be routine. Both Bivariate and multivariate techniques can be used in the analysis, based on the evaluated contrast. Estimating test accuracy and prediction of outcome require different approaches.
- External (clinical) validation should preferably be based on repeated studies in other, similar populations. Also, systematic reviews and meta-analysis have a role.

Introduction

Although the ultimate objective of the diagnostic phase is to optimise the patient's prognosis by enabling the clinician to choose an adequate therapeutic strategy, an accurate diagnostic assessment is a first and indispensable step in the process of clinical management.

Making a useful clinical diagnosis implies classifying the presented health problem of a patient in the context of accepted nosological knowledge. This diagnostic classification may result in confirmation or exclusion of the presence of a certain disease, in the selection of one disease from a set of possibly present diseases, or in the conclusion that a number of diseases are present simultaneously.[1] Also, it can be concluded that, given present knowledge, a further diagnostic classification than the observed symptomatology cannot be achieved. Sometimes such a classification is not worthwhile, considering the balance between expected gain in certainty, the burden of making a definitive diagnosis, and relevant therapeutic consequences.

Apart from making a diagnostic classification, the diagnostic process may be aimed at assessing the clinical severity or monitoring the clinical course of a diagnosed condition. Another very important clinical application is documenting the precise localisation or shape of a diagnosed lesion to support further, for example surgical, decision making.

A potential new diagnostic test must first go through a phase of pathophysiological and technical development, before its clinical

effectiveness in terms of diagnostic accuracy or prognostic impact can be evaluated. The methodology discussed in this book, focused on clinical effectiveness, is applicable to the further evaluation of tests that have succesfully passed this early development.

A basic question to be anwered, then, is: what is the probability that this particular patient with this particular symptomatology or test result has a certain disorder or a combination of disorders? Obtaining an evidence-based answer, using clinical epidemiological research data, requires an analysis of the association between the presented symptomatology or test result and the appropriate diagnostic classification, that is, the presence or absence of certain diagnoses.

This chapter deals with principles, design and pitfalls of cross-sectional diagnostic research. In this context, cross-sectional research includes studies in which the measured test results and the health status to be diagnosed essentially represent one point in time for each study subject.[2]

Diagnostic research on test accuracy: the basic steps to take

All measures of diagnostic association[3] (Chapters 1 and 7) can be derived from research data on the relation between test results and a reference standard diagnosis. A valid data collection on this relation is the main point of concern,[4] and the various measures can be calculated by applying straightforward analytical methods. Research data for the purpose of diagnostic discrimination are generally collected in cross-sectional research, irrespective of the diagnostic parameters to be used.

As usual in research, a first requirement is to specify the research question. Second, the most appropriate study design to answer this question has to be outlined. A third step is to operationalise the determinants (test(s) to be evaluated, relevant modifiers of diagnostic accuracy, and possible confounding variables) and outcome (generally the presence or absence of the disorder to be diagnosed). Further, the study population, the corresponding inclusion and exclusion criteria, and the most appropriate recruitment procedure have to be further specified. Finally, an adequate data analysis must be planned and achieved.

The research question: contrast to be evaluated

In short, the diagnostic research question should define:

- The test or test set to be evaluated
- The clinical problem for which the use of these test(s) is considered possibly relevant

- Whether the planned study should evaluate (1) the potential of the test procedure to discriminate between subjects with and without a target disorder in an ideal situation of extreme contrast, or (2) to what extent it could be useful in a daily practice clinical setting (where discrimination is, by definition, more difficult).

Box 3.1 The research question

(a) *Contrast to be evaluated*
- single test
- comparing single tests
- additional testing
- comparing diagnostic strategies

(b) *Define the clinical problem*

(c) *Extreme contrast or practice setting*

A key issue is the specific contrast to be evaluated (a). The question can be, for example, What is the discriminative power of one specific test or test procedure to be applied for a certain clinical diagnostic problem (single test)? However, the focus may also be on the discriminative power of a new test compared to the best test(s) already available for a similar clinical diagnostic problem (comparing single tests). For clinical practice, it is important to determine the added value of further (for example more invasive) testing, given the tests already performed (additional testing),[5] or to evaluate the most accurate or efficient diagnostic test set or test sequence (diagnostic strategy) for a certain diagnostic problem. In general, we recommend that in studying a new, more invasive or more expensive test, already available less invasive or less expensive tests with fewer adverse effects should be also included. This makes it possible to critically evaluate the new test's net contribution, if any. Also, it is informative to include the clinician's diagnostic assessment (without knowing the test result) as a separate test. The performance of the new approach can then be evaluated as to its added value compared to the doctor's usual diagnostic performance or "black box".

Regarding the clinical problem studied (b), it is of key importance not only to define the target disorder(s) to be diagnosed, but also the clinical setting and clinical spectrum (for example, early or later in the development of the disorder, and degree of severity) at which one is primarily aiming. It is crucial whether the investigator wants to evaluate the validity of a test for diagnosing a possible disease in its early phase in a primary care setting, or to diagnose more advanced disease in an outpatient clinic or hospital setting, with patients selected by referral based

on suspect symptoms or previous tests.[6,7] This is dealt with in more detail in Chapters 2 and 6.

As to (c), critical appraisal of the state of current knowledge is important for defining an optimally efficient research strategy. For instance, if nothing at all is known yet about the discriminative power of a test, it is more efficient – in terms of reducing the burden for study patients, the sample size, the resources, and the time needed for the study – first to evaluate whether the test discriminates between clearly diseased and clearly non-diseased subjects. If the test does not discriminate in such a Phase I study (Feinstein[2], Sackett and Haynes, Chapter 2 of this book) any further, usually larger and longer, studies evaluating a more difficult contrast between clinically similar study subjects will be useless: the index test cannot be expected to add anything valuable to clinical practice anyhow.

The specification of these three aspects of the research question is decisive for designing the optimal study methodology. Aspect (c) was extensively addressed in Chapter 2.

Outline of the study design

Because study questions on diagnostic accuracy generally evaluate the association between (combinations of) test results and health status (mostly the presence or absence of a target disorder), a cross-sectional design is a natural basic design option. However, this basic design has various modifications, each with specific pros and cons in terms of scientific requirements, burden for the study subjects, and efficient use of resources.

Box 3.2 Study design

General approach
- survey of total study population
- case–referent approach
- test based enrolment

Direction of data collection
- prospective
- ambispective
- retrospective

General approach

The most straightforward approach of the cross-sectional design is a survey of the study population to determine the test distribution and the presence of the target disorder simultaneously. Examples are a survey on the relationship

between intermittent claudication and peripheral arterial occlusive disease in an elderly population,[8] and a study in a consecutive series of sciatica patients to determine the accuracy of history and clinical examination.[9]

Another option is the case–referent approach, starting from an already known disease status (for example present or absent) as the criterion for enrolment, the test result being determined afterwards in the study patients. This design type may be more efficient or more acceptable when the disease under study is infrequent or when the reference standard procedure is highly invasive to the patient, for example in pancreatic cancer, or very expensive.

A further approach is test based enrolment, where the available test result (such as positive or negative) is the criterion for recruitment, with the disease status being determined afterwards. This modification may be preferable when test results are easily available from routine health care. An example regarding the latter is a study on the diagnostic value of "fatigue" for detecting anaemia in primary care, comparing patients presenting with fatigue and a control group as to haematological parameters.[10]

In the context of the cross-sectional design, efficient sampling of the studied distributions may be artificially facilitated at the determinant (test) or the outcome (target disorder) side. For example, in order to achieve a balanced data collection over the relevant categories, a stratified sample can be drawn from the various test result levels or from various parts of the whole disease spectrum, from clearly no disease to most severe disease. Also, the contrast between the categories of the diseased and non-diseased subjects can be enhanced by limiting the sampling to those who have been proved to be clearly diseased and those proved to be completely healthy. The latter approach is applied in planning a Phase I study (Chapter 2), which is essentially a case–referent study. Because of a sharp contrast in disease spectrum between diseased and non-diseased, sensitivity and specificity will be optimal, and as by sampling of cases and referents the "prevalence" in the study population can also be artificially optimised (with, for example, a disease prevalence of 50%), Phase I studies generally need a relatively small number of subjects to be included. Moreover, for a Phase I study the subjects in the "case group" (the diseased) and those in the reference group (the healthy or non-diseased subjects) can be specifically selected from populations consisting of subjects who have already undergone a "reference standard" procedure with maximum certainty.

Direction of data collection

Whereas Phase I and Phase II studies (according to Sackett and Haynes, Chapter 2) may be based on either retrospective or prospective identification of subjects with a certain diagnosis or test status, Phase III studies must usually be prospectively planned. The latter start from a study

population of subjects comparable with those who will be tested in clinical practice. In such studies it is not known in advance who is diseased and who is not, and the clinical characteristics of the two are therefore very similar (which in fact is the reason that testing is clinically necessary at all). Because the clinical contrast is much less pronounced, and as the prevalence of diseased subjects is usually much lower than 50%, substantially larger sample sizes are generally needed than in Phase I studies.

Also, when the subject selection is prospective the data collection can be partly retrospective (ambispective approach). For instance, if patient history is an important element of the diagnostic test to be evaluated (such as when studying the diagnostic value of rectal bleeding, palpitations, or psychiatric symptoms in the preceding 6 months), information about the past is included in the test result. Essential, however, is that the test result status, albeit based on historical information, is evaluated and interpreted as to its diagnostic accuracy at the very moment that the patient "history" is taken.

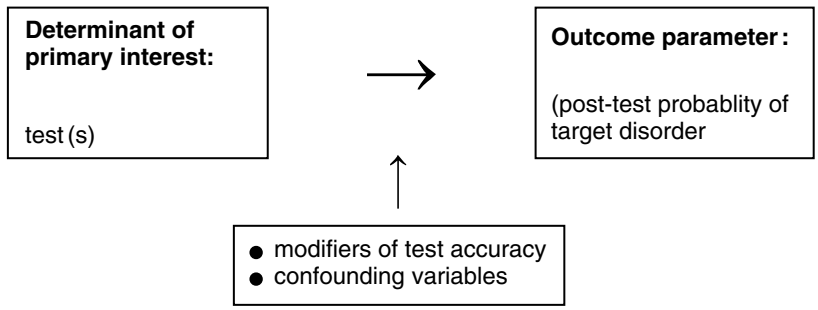
The "direction" of the sampling and the data collection must be decided upon in advance. In addition, and secondary to scientific considerations, practical issues may play a role, such as the availability of data and the efficiency of its collection. Prospectively planned data collections often take more time but are generally more valid, as the procedure and the quality of the data collection can be optimised beforehand. But this is not necessarily always the case. Valid data may be already available in a well documented database of an appropriate study population with an adequately described epidemiological (morbidity) numerator and (population) denominator, and with all relevant covariables present. Especially when the clinical indication to perform the test is appropriately defined (for example coronary angiography in instable angina pectoris) and recorded, and when all eligible patients can be assumed to be included, this is an option. Moreover, a prospective data collection may sometimes imply a higher risk of bias than a retrospective approach. For example, if participating clinicians know that they are being observed in a study of the accuracy of their usual diagnostic assessment compared to an independent standard or panel, their behaviour will be easily be influenced in the context of a prospective design (Hawthorne effect). However, in a retrospective design the availability of complete and well standardised data and the controlling of the subject selection process are often problematic.

Operationalising determinants and outcome

Determinants

As in any (clinical) epidemiological study, research questions on diagnostic accuracy can be operationalised in a central "occurrence relation"[11] between independent and dependent variables.

Box 3.3 The occurrence relation in diagnostic research



The independent variable or determinant of primary interest is the test result to be evaluated, and the primary dependent or outcome variable is (presence or absence of) the target disorder. When evaluating a single test, the test results in all study subjects are related to the reference standard. In fact, we are then comparing testing (yielding the post-test probabilities of the disorder D, for example for positive and negative test results) with not testing (expressed in the pretest probability of D). When two or more tests are compared, we have a number of separate determinants that are contrasted as to their discriminatory power. In studying the value of an additional (for example more invasive) test, given the tests already performed, the discrimination of applying all other tests is compared with that of applying all other tests plus the additional one. And to evaluate the most accurate or efficient diagnostic test set or strategy for a certain clinical problem, the performances of all the considered test combinations and sequences must be compared. To be able to make these comparisons, all separate tests have to be performed in all study subjects.

The accuracy of diagnostic tests may vary in relation to subject characteristics, such as gender, age, and comorbidity. For example, in studying the diagnostic accuracy of mammography in the detection of breast cancer, it is important to consider that this accuracy depends on age, gender, and the possible presence of fibroadenomatosis of the breasts. To evaluate the influence of such modifiers of test accuracy, these have to be measured and included in the analysis. In fact, we are dealing here with various subgroups where the diagnostic accuracy may be different. Effect modifying variables can be accounted for later in the analysis, for example by stratified analysis (subgroup analysis) of the measures of diagnostic association, or by introducing interaction terms in the logistic regression analysis (Chapter 7).[12,13] Because diagnostic assessment can be seen as optimal discrimination between subgroups with a different probability of

disease, effect modifying variables can also be considered as additional diagnostic tests themselves.

Confounding variables are independent extraneous determinants of the outcome that may obscure or inflate an association between the test and the disorder. They are essentially related to both the test result and the outcome. For example, in studying whether the symptom fatigue is predictive for a low blood haemoglobin level, it is important to know which study subjects have previously taken oral iron, as this can improve the fatigue symptoms and enhance the haemoglobin level as well. A confounder can only be controlled for if it is considered beforehand and measured, which requires insight into relevant external influences. In diagnostic research the term "confounding variable" is used in a different, more pragmatic sense than in aetiologic research, as consistent diagnostic correlations do not need to be fully causally understood to be useful.

Generally, according to Bayes' theorem, the pretest probability of the target disorder is seen as a basic determinant of the post-test probability, independent of the accuracy of the applied tests. However, the clinical spectrum of the disorder may be essentially different in high and low prevalence situations. Because the clinical spectrum can influence test accuracy (Chapters 1, 2, and 6), it is then crucial to measure separately spectrum characteristics, such as disease severity, and frequency of occurrence as such. Spectrum characteristics can then be analysed as modifiers of test accuracy.

Good test reproducibility is a requirement for good accuracy in practice. Therefore, when the test under study is sensitive to inter- or intraobserver variability, documentation and, if possible, reduction of this variability is important. Documentation can be achieved in a pilot study or in the context of the main study. For example, in a study of the accuracy and reproducibility of erythrocyte sedimentation rate (ESR) measurements in general practice centres, for measuring an identical specimen a clinically relevant range between practices from 4 to 40 mm/1 h was observed. The average coefficient of variation (CV: standard deviation as a percentage of the mean) was 37% between practices and 28% within practices.[9] Observer variability can be reduced by training. In the same ESR study, the average inter- and intrapractice CVs were reduced by training to 17% and 7%, respectively. The accuracy of a test can be evaluated in relation to the achieved reproducibility. This reproducibility must then, for practical purposes, be judged as to its clinical acceptability and feasibility.

Outcome: the reference standard

Principles

Establishing a final and "gold standard" diagnosis of the target disorder is generally more invasive and expensive than applying the studied diagnostic

test. It is exactly for this reason that good test accuracy (for example a very high sensitivity and specificity) would be very useful in clinical practice to make a satisfactory diagnostic assessment without having to perform the reference standard. However, in performing diagnostic research the central outcome variable – the presence or absence of the target disorder – must be measured, as it is the reference standard for estimating the test accuracy. A real gold – that is, perfect – standard test, with 100% sensitivity and specificity, is exceptional. Even histological classification and MRI imaging are not infallible, and may yield false positive, false negative and uninterpretable conclusions. Therefore, the term "reference standard" is nowadays considered better than "gold standard".

Box 3.4 The reference standard

Principles
- to be applied on all included subjects
- independent assessment of test and standard
- standardised protocol

Possible problems with the reference standard
- imperfect: classification errors
- pathophysiological concept not well defined (independent from clinical presentation)
- incorporation bias
- too invasive
- too complex

Possible solutions
- pragmatic criteria
- independent expert panel
- clinical follow up: delayed-type cross-sectional study
- tailormade standard protocol
- prognostic criterion

The reference standard to establish the final diagnosis (outcome) should be applied for all included subjects. Applying different standard procedures for different patients may yield an inconsistent reference for the evaluated test, as each of the "standards" will have its own idiosyncratic error rate.

The results of the test for each patient should be interpreted without knowledge of the reference standard results. Similarly, the reference standard result should be established without knowing the outcome of the test under study. Where such blinding is not maintained, "test review bias"

and "diagnosis review bias" may occur: non-independent assessment of test and reference standard, mostly resulting in overestimation of test accuracy.

The reference standard must be properly performed and interpreted using standardised criteria. This is especially important when the standard diagnosis depends on subjective interpretations, for example by a psychiatrist, a pathologist, or a radiologist. In such cases inter- and even intraobserver variability in establishing the standard can occur. For example, in evaluating the intraobserver variability of MRI assessment as the standard for nerve root compression in sciatica patients, the same radiologist repeatedly scored the presence of root compression as such consistently (κ: 1.0) but the site of root compression only moderately (κ: 0.60).[14] In these situations, training sessions and permanent documentation of performance are important.

Problems and solutions

Apart from the limitations in reaching a 100% perfect standard diagnosis, meeting the requirements for a reference standard can be problematic in various ways.

For many conditions a reference standard cannot be measured on the basis of a well defined pathophysiological concept, independent of the clinical presentation. Examples are sinusitis, migraine, depression, irritable bowel syndrome, and benign prostatic hyperplasia. When, in such cases, information related to the test result (for example symptom status) is incorporated into the diagnostic criteria, "incorporation bias" may result in overestimation of test accuracy. Furthermore, a defined reference standard procedure may sometimes be too invasive for research purposes. For instance, when validating urinary flow measurement it would be unacceptable to apply invasive urodynamic studies to those with normal flow results.[15] And in studying the diagnostic value of non-acute abdominal pain in primary care for diagnosing intra-abdominal cancer, one cannot imagine that all patients presenting with non-acute abdominal pain would be routinely offered laparotomy.[16] In addition, one may doubt whether such laparotomy, if it were to be performed, could always provide an accurate final diagnosis in the very early stage of a malignancy. Another problem can be that a complex reference standard might include a large number of laboratory tests, so that many false positive test results could occur by chance.

For these problems, practical solutions have been sought.

Pragmatic criteria

The absence of a well defined pathophysiological concept can sometimes be overcome by defining consensus based pragmatic criteria. However, if applying such reference standard criteria (such as a cut-off value on a

depression questionnaire) is no more difficult than applying the test under study, evaluating and introducing the test will not be very useful.

Independent expert panel

Another method is the composition of an independent expert panel that, given general criteria and decision rules for clinical agreement, can assign a final diagnosis to each patient, based on the available clinical information. To achieve a reasonably consistent classification it is important that the panel is well prepared for its task, and a training session or pilot study using patients with an already established diagnosis is recommended. The agreement of the primary assessments of the individual panel members, prior to reaching to consensus, can be documented.

Clinical follow up: delayed-type cross-sectional study

When applying a definitive reference standard is too invasive or otherwise inapplicable at the very moment that the test should be predictive for the presence of the target disorder, a good alternative can be follow up of the clinical course during a suitable predefined period. Most diseases that are not self-limiting, such as cancers and chronic degenerative diseases, will usually become clinically manifest a period of months or a year or so after the first diagnostic suspicion (generally the moment of enrolment in the study) was raised. The follow up period should not be too short, in order to give early phase disorders the chance to become manifest and therefore to have a minimum number of "false negative" final diagnoses. Nor should it be too long, so as to avoid the final diagnosis after follow up being related to a new disease episode started after the baseline "cross-section" (false positives).[16,17] In addition, it would be ideal to collect the follow up data that are decisive for making the final diagnosis independently from and blinded to the health status and test results at time zero, and also to blind the final outcome assessment for these test results.

One should take into account that "confounding by indication" can occur: management decisions during follow up are possibly related to the health status at baseline, and might therefore influence the clinical course and the probability of detecting the target disorder. This is especially a point of concern in studying target disorders with a rather variable clinical course, becoming clinically manifest dependent on management decisions.

In contrast to what is commonly thought, it has to be acknowledged that the described method of clinical follow up should not be considered as a "follow up" or "cohort" study, as the focus is not on relating the time zero data to a subsequently developing (incident) disorder. In fact, the follow up procedure is aimed at retrospectively assessing the (prevalent) health status at time zero, as a substitute for establishing the reference standard

diagnosis of the target disorder immediately at time zero itself. Therefore, this design modification can be designated a "delayed-type cross-sectional study", instead of a follow up study.[18]

The expert panel and the clinical follow up can be combined in a composite reference standard procedure, in which the outcome after follow up is evaluated and established by the panel.[15]

Tailormade standard protocol

In some situations, for example in diagnostic research on psychiatric illnesses, it might be difficult to separate test data at time zero (for example the presence of anxiety) from the information needed to make a final assessment (anxiety disorder). As mentioned earlier, in such situations incorporation bias may be the result. If test data are indeed an essential part of the diagnostic criteria, one cannot avoid balancing a certain risk of incorporation bias against not being able to perform diagnostic research at all, or making a final diagnosis while ignoring an important element of the criteria. Often, one can find a practical compromise in considering that for clinical purposes it is sufficient to know to what extent the available diagnostic tests at time zero are able to predict the target disorder's becoming clinically manifest during a reasonably chosen follow up period.[17] There is also the option to ask the expert panel to establish the final diagnosis first without the baseline test data, and then to repeat it with these data incorporated. This can be done while adding an extra blinding step, such as randomly rearranging the order of anonymised patient records. If there then appear to be important differences in the research conclusions, this should be transparently reported and discussed as to the clinical implications.

When it is impossible to meet the principle that the reference standard should be similarly applied to all study subjects irrespective of their health or test result status, "next best" solutions can be considered. For example, to determine the accuracy of the exercise electrocardiogram (ECG) in primary care settings, it might be considered medically and ethically unjustified to submit those with a negative test result to coronary angiography. For these test negatives a well standardised clinical follow up protocol (delayed-type cross-sectional study) might be acceptable. This option is particularly important when the focus is on exercise ECG in patients who have a relatively low prior probability of coronary heart disease. For this spectrum of patients, results of a study limited to those who would be clinically selected for coronary angiography would be clearly not applicable.[19–21] If one would still prefer an identical standard procedure for primary care patients, irrespective of previous test results, one could also consider submitting all patients to the clinical follow up standard procedure. In order to have some validation of this procedure, for the

subgroup who had an angiography one can compare the standard diagnoses based on the follow up with the angiography results.

In summary, although a completely identical and "hard" reference standard for all included study subjects is the general methodological paradigm, this is not always achievable. In such situations, given the described limitations and the suggested alternative approaches, establishing a well documented and reproducible reference standard protocol – indicating the optimal procedure for each type of patient – may be not only the best one can get but also sufficient for clinical purposes.

Prognostic criterion

Diagnostic testing should ultimately be in favour of the patient's health, rather than just an assessment of the probability of the presence of disease. In view of this, the reference standard procedure can sometimes incorporate prognosis or consequences for clinical management.[22,23] It is then a starting point for further decision making (for example, whether treatment is useful or not) rather than a diagnostic endpoint. This is especially relevant in situations where an exhaustive nosological classification is less important, and when management is based primarily on the clinical assessment (for example in deciding about physiotherapy in low back pain, or referral to a neurosurgeon in sciatica[14]). Sometimes making a final diagnosis is less important than a prognosis, in view of the clinical consequences (incidental fever) or the lack of a solid diagnostic consensus (the pyriformis syndrome). In establishing a "prognostic reference criterion", a pitfall is that prognosis can be influenced by interfering treatments. In this context, methods for unbiased assessment of the impact of testing on patient prognosis are important (Chapter 4). It is to be expected that with the progress of DNA testing for disease, prognostic relevance will increasingly be the reference standard.

Standard shift as a result of new insights

At certain moments during the progress of pathophysiological knowledge on diagnosis, new diagnostic tests may be developed that are better than the currently prevailing reference standard. However, if this possibility is systematically ignored by reducing diagnostic accuracy research to just comparing new tests with traditional standards, possible new reference standards would never be recognised, as they would always seem less accurate than the traditional ones. Therefore, pathophysiological expertise should be involved in the evaluation of diagnostic accuracy. Examples of a shift in reference standard are the replacement of the clinical definition of tuberculosis by the identification of *Mycobacterium tuberculosis*, and of old imaging techniques by new ones (see also Chapter 1).

Specifying the study population

As in all clinical research, the study population for diagnostic research should be appropriately chosen, defined, and recruited. The selection of patients is crucial for the study outcome and its external (clinical) validity. For example, as has already been emphasised, it is widely recognised that diagnostic accuracy is very much dependent on the spectrum of included patients and the results of relevant tests performed earlier, and may differ for primary care patients and patients referred to a hospital.[6,7,21]

Given that the test has successfully passed Phase I and Phase II studies (Chapter 2), the starting point is the clinical problem for which the test under study should be evaluated as to its diagnostic accuracy, taking the relevant healthcare setting into account. For example, the study can address the diagnostic accuracy of clinical tests for sciatica in general practice, the accuracy of ECG recording in outpatients with palpitations without a compelling clinical reason for immediate referral, or the diagnostic accuracy of the MRI scan in diagnosing intracerebral pathology in an academic neurological centre. The study population should be representative for the "indicated", "candidate", or "intended" patient population, also called the target population, thereby being clinically similar to the group of patients in whom the validated test is to be applied in practice.[24,25] The "intention to diagnose" should be the key criterion for the study of presented clinical problems. For the evaluation of population screening of asymptomatic subjects, such as in the context of breast cancer screening or hypertension case finding, a study population similar to the target population "intended to be screened" is required.

Box 3.5 Study population

- In accordance with studied clinical problem
- "Intention to diagnose" or "intention to screen"
- Inclusion criteria corresponding with "indicated" population
- Consecutive series ("iatrotropic") or target population survey

The next step is generally straightforward: the specification of the relevant process of selection of patients for the study (preferably corresponding with the clinical problem and the healthcare setting requirements), and the relevant inclusion citeria (in accordance with the indicated population and the relevant patient spectrum). Exclusion criteria may also be defined, for example identifying those patients for whom the

reference standard procedure is too risky or too burdensome. The importance of explicitly formulated entry criteria is demonstrated by a study on the diagnostic value of reported collapse for the diagnosis of clinically relevant arrhythmias in general practice: if the inclusion was based on presented symptoms the odds ratio (OR) was 1.9, whereas for an inclusion based on coincidentally finding a pulse < 60 or > 100 bpm or an irregular pulse, the OR was 10.1.[26] Investigators should include "indicated" patients as long as there is no compelling reason not to do so, in order that after the study a non-evidence-based testing practice will not be introduced or maintained for relevant parts of the "real life" patient spectrum. In this context it is emphasised that, for example in the elderly, comorbidity in addition to the possible presence of the target disorder is often an important aspect of clinical reality.[27] For the clinical applicability of the study measuring comorbidity and studying it as a modifier of diagnostic accuracy is the preferred approach, instead of excluding it.

In the section on study design we discussed the choice between population survey and disorder- or test-oriented subject selection, covering the principal starting point of patient recruitment. In addition, the pros and cons of the various options for practical patient recruitment should be considered. When problems presented to clinicians are studied, recruiting (a random sample of) a series of consecutively presenting patients who meet the criteria of the indicated population is most sensible for clinical validity purposes. This should preferably be supported by a description of the patient flow in health care prior to presentation. Sometimes, however, it may take too long to await the enrolment of a sufficiently large consecutive series – for example when the clinical problem or target disorder is very rare, or when a useful contribution to rapidly progressing diagnostic knowledge can only be made within a limited period. In such situations, active surveys of the target population or sampling from a patient register can be alternatives. However, it should be borne in mind that such methods may yield a population of study subjects with a different clinical spectrum. Also, for such subjects the indication to test is less clear than for patients who experienced an "iatrotropic" stimulus[2] to visit a doctor at the very moment that they want their health problem to be solved.

Of course, for validating test procedures to be used in population screening, an active survey of a study population similar to the target population is the best approach.

In order to enhance external validity, it is essential to add key demographic and clinical characteristics of the identified study population, and to evaluate non-response in relation to these characteristics. Furthermore, all important steps in the study protocol, with data on specific subgroups, including subgroup non-response, should be documented. This can be supported by a flow diagram.

Adverse effects of test and reference standard

Apart from its accuracy, the performance of a test has to be evaluated as to its (dis)comfort to both patient and doctor. In particular, a test should be minimally invasive and have a minimal risk of adverse effects and serious complications. Measuring these aspects in the context of a diagnostic accuracy study can add to the comparison with other tests as to their clinical pros and cons.

For the research community, it is also important to learn about the invasiveness and risks of the reference standard used. For example, if in the evaluation of the positive test results of Haemoccult screening colonoscopy, sigmoidoscopy or double contrast barium enema were to be used, one might expect complications (perforation or haemorrhage) once in 300–900 subjects investigated.[28] Researchers can use the experience reported by colleagues studying similar problems, in order to make an optimal choice of reference standard procedure, taking possible adverse events into consideration.

Statistical aspects

Box 3.6 Statistical aspects

- Sample size
- Bivariate and multivariable analysis
- Test accuracy or prediction of outcome
- Single test; comparing tests or strategies; additional testing
- Difference between diagnostic and aetiologic data analysis

In the planning phase of the study the required sample size should be considered. To evaluate the relationship between a dichotomous test and the presence of a disorder, one can use the usual programs for sample size estimation. For example, for a case–referent study with equal group sizes, accepting certain values for type I and type II errors (for example 0.05 and 0.20, respectively) and using two-sided testing, one can calculate the number of subjects needed per group to detect a minimum sensitivity (proportion of test positives among the diseased, for example at least 0.60) assuming a certain maximum proportion of test positives among non-diseased (for example 0.20, implying a specificity of at least 0.80). For the example above the calculation using the program EPI-Info[29] would yield a required number of 27 cases and 27 referents. Of course, when performing a cross-sectional study prospectively in a consecutive series with a low

expected prevalence of the target disorder (unequal group sizes), the required sample will be much higher. Also, if a number of determinants is simultaneously included in the analysis, the required sample size is higher: as a rule of thumb, for each determinant at least 10 subjects with the target disorder are needed.[30]

Data analysis in diagnostic research follows the general lines of that in clinical epidemiological research. For single tests the first step is a Bivariate analysis focused on one predictive variable only, for example in a 2×2 table in the case of a dichotomous test. It is possible to stratify for modifiers of accuracy, thereby distinguishing relevant clinical subgroups, and to adjust for potential confounding variables. Point estimates and confidence intervals for the measures of diagnostic accuracy can be determined. Subsequently, there are various options for multivariable analysis, taking the influence of multiple independent variables into account simultaneously. Multiple logistic regression is especially useful for analysing accuracy data.[12,13] These data analytical challenges in diagnostic research are discussed in detail in Chapter 7.

It is important to distinguish the analytical approach focusing on the accuracy of individual tests from the analysis where an optimal prediction of the presence of the studied disorder in patients is at stake. In the first, the dependent variable may even be test accuracy itself, as a function of various determinants. In the latter, a diagnostic prediction model can be derived with disease probability as the dependent variable, and with various tests, demographic, and clinical covariables as independent variables.[18,31]

When a number of tests are applied there are various analytical options. First, the accuracy of all tests can be determined and compared. Furthermore, using multivariate analysis such as multiple logistic regression, the combined predictive power of sets of test variables can be determined. Moreover, starting from the least invasive and most easily available test (such as history taking), it can be evaluated whether adding more invasive or more expensive tests contributes to the diagnosis. For example, the subsequent contributions of history, physical examination, laboratory testing, and more elaborate additional investigations can be analysed, supported by displaying the ROC curves (with areas under the curve) of the respectively extended test sets (see Chapter 7).[8,26]

It must be acknowledged that data analysis in diagnostic research is essentially different from aetiologic data analysis. The principal difference is that aetiologic analysis usually focuses on the effect of a hypothesised aetiologic factor adjusted for the influence of possible confounders, thereby aiming at a causal interpretation. In diagnostic research the focus is on identifying the best correlates of the target disorder irrespective of any causal interpretations. It is sufficient if these correlates (tests) can be systematically and reproducibly used for diagnostic prediction. Whereas in

aetiologic analysis there is a natural hierarchical relation between the possible aetiologic factor of interest and the covariables to be adjusted for, such a hierarchy is absent for the possible predictors in diagnostic research. This implies that diagnostic data analysis can be more pragmatic, seeking for the best correlates.

External validation

Analyses of diagnostic accuracy in the collected data set, especially the results of multivariable analyses, may produce too optimistic results that may not be reproducible in clinical practice or similar study populations.[32] Therefore, it is advisable to perform one or more separate external validation studies in independent but clinically similar populations.

Box 3.7 External (clinical) validation

- Results based on study data may be too optimistic
- "Split-half" analysis is no external validation
- Repeated studies in other, similar populations are preferred
- First exploration: compare first included half with second half
- Role of systematic reviews and meta-analysis

Sometimes authors derive a diagnostic model in a random half of the research data set and test its performance in the other half (split-half analysis). This approach is not addressing the issue of external validation: in fact, it only evaluates the degree of random error at the cost of possibly increasing such error by reducing the available sample size by 50%.[18] Also, other methods using one and the same database do not provide a real external validation. An exploratory approximation, however, could be to compare the performance of the diagnostic model in the chronologically first enrolled half of the patients, with that in the second half. The justification is that the second half is not a random sample of the total, but rather a subsequent clinically similar study population. However, totally independent studies in other, clinically similar settings will be more convincing. In fact, over time, various studies can be done in comparable settings, enabling diagnostic systematic reviews and meta-analyses to be performed. This may yield a constantly increasing insight into the performance of the studied diagnostic test, both in general and in relevant clinical subgroups (Chapter 8).

References

1 van den Akker M, Buntinx F, Metsemakers JF, Roos S, Knottnerus JA. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *J Clin Epidemiol* 1998;**51**:367–75.

2 Feinstein AR. *Clinical Epidemiology. The architecture of clinical research*. Philadelphia: WB Saunders, 1985.

3 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology. A basic science for clinical medicine*. Boston: Little, Brown and Company, 1991.

4 Lijmer JG, Mol BW, Heisterkamp S *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.

5 Moons KGM, van Es GA, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes's theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1997;**8**:12–17.

6 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.

7 Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;**45**:1143–54.

8 Stoffers HEJH. *Peripheral arterial occlusive disease. Prevalence and diagnostic management in general practice*. Thesis. Maastricht: Datawyse, 1995.

9 Dinant GJ, Knottnerus JA, van Aubel PGJ, van Wersch JWJ. Reliability of the erythrocyte sedimentation rate in general practice. *Scand J Prim Health Care* 1989;**7**:231–5.

10 Knottnerus JA, Knipschild PG, van Wersch JWJ, Sijstermanns AHJ. Unexplained fatigue and hemoglobin, a primary care study. *Can Fam Phys* 1986;**32**:1601–4.

11 Miettinen OS. *Theoretical Epidemiology. Principles of occurrence research in medicine*. New York: John Wiley & Sons, 1985.

12 Spiegelhalter DJ, Knill-Jones RD. Statistical and knowledge-based approaches to clinical decision support systems, with an application to gastroenterology. *J Roy Stat Soc* Ser A 1984;**147**:35–76.

13 Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. *Med Decision Making* 1992;**12**:93–108.

14 Vroomen PCAJ. *The diagnosis and conservative treatment of sciatica*. Maastricht: Datawyse, 1998.

15 Wolfs GGMC. *Obstructive micturition problems in elderly males, prevalence and diagnosis in general practice*. Maastricht: Datawyse, 1997.

16 Muris JW, Starmans R. *Non acute abdominal complaints. Diagnostic studies in general practice and outpatient clinic*. Thesis. Maastricht, 1993.

17 Warndorff DK, Knottnerus JA, Huijnen LG, Starmans R. How well do general practitioners manage dyspepsia? *J Roy Coll Gen Pract* 1989;**39**:499–502.

18 Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. *Primary Care* 1995;**22**:341–63.

19 Green MS. The effect of validation group bias on screening tests for coronary artery disease. *Stat Med* 1985;**4**:53–61.

20 Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. *Med Decision Making* 1987;**7**:139–48.

21 Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;**39**:207–16.

22 Hunink MG. Outcome research and cost-effectiveness analysis in radiology. *Eur Radiol* 1996;**6**:615–20.

23 Moons KGM. *Diagnostic research: theory and application*. Thesis. Rotterdam 1996.

24 Dinant GJ. *Diagnostic value of the erythrocyte sedimentation rate in general practice*. Thesis. Maastricht, 1991.

25 van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Invest Radiol* 1995;**30**:334–40.

26 Zwietering P. *Arrhythmias in general practice, prevalence and clinical diagnosis*. Maastricht: Datawyse, 2000.

27 Schellevis FG, van der Velden J, van de Lisdonk E, van Eijk JT, van Weel C. Comorbidity of chronic diseases in general practice. *J Clin Epidemiol* 1993;**46**:469–73.
28 Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. *Screening for colorectal cancer using the faecal occult blood test, hemoccult*. Cochrane Database Syst Rev 2000;(2):CD001216.
29 Dean AG. *The Epi Info Manual*. Brixton Books, 1996.
30 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87.
31 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;**277**:488–94.
32 Starmans R, Muris JW, Fijten GH, Schouten HJ, Pop P, Knottnerus JA. The diagnostic value of scoring models for organic and non-organic gastrointestinal disease, including the irritable-bowel syndrome. *Med Decision Making* 1994;**14**:208–16.

4 Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research

JEROEN G LIJMER, PATRICK M BOSSUYT

Summary box

- From the patient's perspective the contribution of diagnostic tests to a better health outcome is of primary importance, rather than their correspondence with the "truth".
- Diagnostic test evaluations should therefore focus on the likelihood that tests detect clinical events of interest and the effect that tests can have on those events by way in which the results affect subsequent management decisions.
- Randomised controlled trials of diagnostic tests are feasible, and several designs are possible.
- Each trial design option has its own advantages and disadvantages. Depending on the clinical problem under study, the type of information needed, and the costs of tests or follow up, one design can be preferred over another.
- Whereas in most randomised controlled trials of diagnostic tests so far the point of randomisation coincided with the clinical decision as to whether or not to perform the test, the trial design can be made more efficient by randomising only patients with the test result of interest.
- The randomised trial design can be elaborated both for evaluating the (additional) prognostic value of a single test, and for comparing different test strategies.

- A randomised controlled trial of diagnostic tests should incorporate a prespecified link between test and treatment options to ascertain validity and generalisability.
- Methods to preserve allocation concealment and blinding deserve special attention.
- Sample size calculations need special attention, and must include an estimation of the discordance rate.

Why bother about the prognostic impact of a diagnostic test?

For scientific purposes it is worth knowing whether or not a result from a medical test corresponds to the truth. Can this value be trusted? Is this truly a sign of disease? These are the first questions that come to the mind in the evaluation of medical tests.

From a patient perspective, mere knowledge about the present, true state of things is in most cases not enough. In relieving health problems, information in itself will not suffice. Patients will only benefit from diagnostic tests if the information generated by those tests is correctly used in subsequent decisions on action to restore or maintain the patient's health.

There are several ways in which medical tests can affect a patient's health. First, undergoing the test itself can have an impact. The adverse effects range from slight discomfort and temporary unpleasantness to lasting side effects or death. On the other hand, undergoing an elaborate procedure can also have a non-specific positive effect on patient complaints, regardless of the information that results from it. This can be called the "placebo" effect of testing, and we know very little about its magnitude and modifying factors.

In addition to the effects of the diagnostic procedure itself, the information generated by the test also influences patients. Information on the likely cause of one's health problems or other aspects of health status can have both a positive and a negative effect, albeit limited. As patients, we want to be informed about the origin of our complaints, even in the absence of a cure. Such information may enable us to find better ways of handling them, by developing strategies to limit their disabling impact on our daily activities.

In these cases it is not just the present state of health that is of interest, but also the future course of disease. It follows, then, that the value of information from diagnostic tests lies not only in the past (where did this come from?) or the present (how is it?), but also in the future. Hence, the relevance of diagnostic information is closely related to prognosis: the implications for the future course of the patient's condition.

The first section of this chapter discusses the evaluation of a single test, starting from an evaluation of its prognostic value and then moving on to the consequences for treatment. It closes with a presentation of randomised designs for evaluating test–treatment combinations. The second section contains an elaboration of the methods for comparing and evaluating multiple test strategies, also including randomised clinical trials. The chapter ends with a discussion on practical issues.

How to measure the prognostic impact of a test

A recent example of the assessment of the prognostic value of a test can be found in the literature on the management of carotid disease. Several studies have examined the need to perform duplex ultrasonography in patients with a cervical bruit without further symptoms of cerebrovascular disease. To answer this question, an assessment has to be made of the value of duplex ultrasonography. Such an evaluation will often look at the amount of agreement between the index test (duplex ultrasonography) and the reference test (the best available method to reveal the true condition of the carotid arteries). In this case, the reference test will mostly likely be conventional angiography. If properly conducted, a 2×2 table can be constructed after the study is done and all indicators of diagnostic accuracy can be calculated. Unfortunately, many of the evaluation studies in diagnostic techniques for carotid stenosis performed so far did not meet the design requirements for an unbiased and useful evaluation.[1]

From a patient perspective, one could successfully argue that it is not so much the correspondence with the "truth" that should be of concern, especially not in asymptomatic patients. For these patients, the true value of the information should come from the strength of the association between data on the presence and severity of carotid stenosis and the likelihood of vascular events in the near future. The appropriate reference standard for such an evaluation will not be a diagnostic procedure. Instead, one should look for clinical information collected through a meticulous follow up of all patients subjected to the index test.

Figure 4.1 illustrates the general design of such a study. All patients with cervical bruits without previous cerebrovascular disease are eligible for the

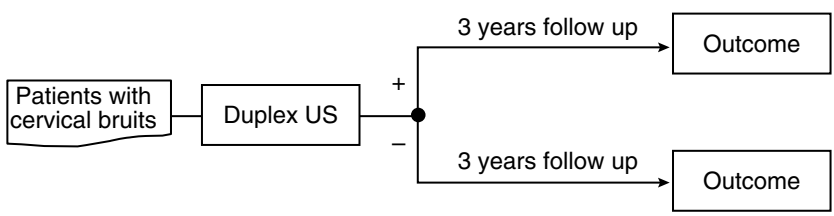


Figure 4.1 Prognostic study.

Table 4.1 Prognostic value of duplex ultrasonography.

	Poor outcome	Favourable outcome	
Stenosis $\geq 80\%$	63 (47%)	72 (53%)	135
Stenosis $< 80\%$	113 (20%)	451 (80%)	564
Total	176	523	699

study. A duplex ultrasonogram of the right and left common and internal carotid arteries is performed in all patients and the percentage of stenosis measured. Ideally, none of the patients receives treatment. Subsequently patients are followed by regular outpatient visits and telephone interviews. The following clinical indicators of poor outcome are recorded: TIA (transient ischaemic attack), stroke, myocardial infarction, unstable angina, vascular deaths, and other deaths.

With data recorded in such a study standard diagnostic accuracy measures can be calculated to express the prognostic value of a test. Table 4.1, based on data published by Lewis et al.,[2] shows positive and negative predictive values of 47% and 80%, respectively, in predicting a poor outcome for a stenosis $\geq 80\%$, as detected on duplex ultrasound. The data also showed that the relative risk of a stenosis $\geq 50\%$ for a TIA or stroke was 2.3. However, insufficient data were presented to reconstruct the 2×2 table for this cut-off point.

The study in Figure 4.1 can provide an answer to the question whether or not a test is able to discriminate between different risk categories for a specific event. Such prognostic information, although of value to patients and healthcare professionals, does not answer the question as to whether there is an intervention that can improve the prognosis of these patients. To respond to this it is necessary to compare the prognosis for different treatment strategies.

Randomised designs for a single test

A slight modification of the design in Figure 4.1 allows us to measure the prognostic value of a test within the context of subsequent clinical decision making. Instead of treating all patients identically one can randomly allocate them to one of the two treatment strategies, establishing the prognostic value of the test in each arm in a way that is similar to the previous example.

A straightforward comparison of patient outcome in the two treatment arms provides an answer as to which treatment is the most effective for all patients included in the trial. Moreover, an analysis stratified by test result offers the possibility of comparing the effectiveness of the treatment options for groups with identical test results.

This type of design and analysis can be illustrated with another example from the field of cerebrovascular disease. In the management of acute stroke the role of intravenous anticoagulation and duplex ultrasonography of the carotid arteries is unclear. A large trial has been performed with as its primary objective the documentation of the efficacy of unfractionated heparin in the treatment of acute stroke. A secondary objective was an evaluation of the role of duplex ultrasonography in selecting patients for anticoagulation.[3,4] A simplified version of the design of this trial is outlined in Figure 4.2. Patients with evidence of an ischaemic stroke, with symptoms present for more than 1 hour but less than 24 hours, were eligible for the study. A duplex ultrasonogram of the right and left common and internal carotid arteries was performed in all included patients. Subsequently, patients were randomised to treatment with unfractionated heparin or placebo, and followed for 3 months. A favourable outcome after stroke was defined as a score of I or II on the Glasgow Coma Scale and a score of 12–20 on the modified Barthel Index.

Tables 4.2(a) and (b) shows the prognostic value of Duplex ultrasonography in each trial arm. An odds ratio can be calculated for each table. These odds ratios can be interpreted as measures of the *natural prognostic value* (Table 4.2(b)) and the *prognostic value with intervention* (Table 4.2(a)), respectively. Another presentation of the same data gives us Tables 4.2(c) and (d), which provide us with information on the treatment effect in both test result categories. We will call the odds ratios of the latter two tables *treatment effect in test normals* (Table 4.2(d)) and *treatment effect in test abnormals* (Table 4.2(c)). When the test discriminates well between patients that benefit from treatment and those that do not, the treatment effect in test abnormals will differ from that in test normals. The ratio of the odds ratios of Tables 4.2(c) and 4.2(d) can therefore be used as a measure of the prognostic impact of the test.

The study in Figure 4.2 provides information on the treatment effect in all test result categories. In practice, it will not always be necessary or ethical to randomise all patients, as uncertainty may exist only for patients with a specific – say, abnormal – test result. This will be the case when there is information available that the prognosis for normal test results is good and

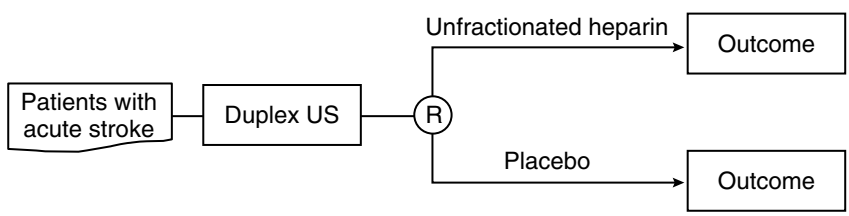


Figure 4.2 Basic RCT of a single diagnostic test.

Table 4.2 Analysis of an RCT of a single diagnostic test.

(a) Unfractionated heparin

	Poor outcome	Favourable outcome	
Stenosis ≥50%	38 (32%)	82 (68%)	120
Stenosis <50%	121 (23%)	400 (77%)	521
Total	159	482	641

(b) Placebo

	Poor outcome	Favourable outcome	
Stenosis ≥50%	51 (47%)	58 (53%)	109
Stenosis <50%	116 (22%)	409 (78%)	525
Total	167	467	634

(c) Stenosis larger than 50% or occlusion

	Poor outcome	Favourable outcome	
Unfract. heparin	38 (32%)	82 (68%)	120
Placebo	51 (47%)	58 (53%)	109
Total	89	140	229

(d) Stenosis smaller than 50%

	Poor outcome	Favourable outcome	
Unfract. heparin	121 (23%)	400 (77%)	521
Placebo	116 (22%)	409 (78%)	525
Total	237	809	1046

(e) Comparison of strategies

	Poor outcome	Favourable outcome	
Duplex US	154 (24%)	491 (76%)	645
No duplex US	167 (26%)	467 (74%)	634
Total	321	958	1279

Duplex US in (e): Decision whether or not to give UFH is based on duplex ultrasonography. The odds ratios and their 95% CI of (a) to (e) are: 1.5 (0.99–2.4), 3.1 (2.0–4.8), 0.53 (0.31–0.90), 1.1 (0.80–1.4), and 0.88 (0.68–1.1). The relative odds ratio of (a)/(b) or (c)/(d) is 0.48.

that patients with such results need no intervention. A logical translation of such a question into a study design would be to randomise only patients with abnormal test results between the different treatment options.

Consider the first example of duplex ultrasonography in patients with cervical bruits. Such a trial could provide evidence that the natural history of patients with a stenosis of less than 50% has a good prognosis. The trial outlined in Figure 4.3 can subsequently answer the question as to whether therapy can improve the prognosis of patients with a stenosis of 50% or more. As in the first example, all patients with cervical bruits without previous cerebrovascular disease are eligible for the study. A duplex ultrasonography of the right and left common and internal carotid arteries is performed in all patients to measure the percentage of stenosis. Subsequently, if the stenosis is 50% or more patients are randomly assigned to receive either aspirin 325 mg a day or placebo. The clinical endpoints – TIA, stroke, myocardial infarction, unstable angina, vascular deaths, and other deaths – are recorded during follow up.

Cote and colleagues performed such a trial in 1995.[5] They randomised 372 neurologically asymptomatic patients with a carotid stenosis of 50% or more between aspirin and placebo. By comparing the outcomes in both treatment arms the effectiveness of treating patients with a stenosis of 50% or more with aspirin was evaluated (treatment effect in test abnormals). In 50 of the 188 patients receiving aspirin and 54 of the 184 receiving placebo a clinical event was measured during follow up, yielding an adjusted hazard ratio (aspirin versus placebo) of 0.99 (95% CI, 0.67–1.46). The authors concluded that aspirin did not have a significant long term protective effect in asymptomatic patients with high grade stenosis (more than 50%).

The trial in Figure 4.3 can also provide information on the accuracy of duplex US in predicting the outcomes of interest (natural prognostic value). This can be done by comparing the outcome in patients in the placebo arm, who all had an abnormal test result, with the outcome in those with a normal test result. A prerequisite for this comparison is that patient management in both of these arms is similar. Table 4.3(a) and (b) shows the crude results and possible comparisons. Note that to calculate the diagnostic accuracy of duplex US it is necessary to correct for the sampling rate of patients with high grade stenosis.[6]

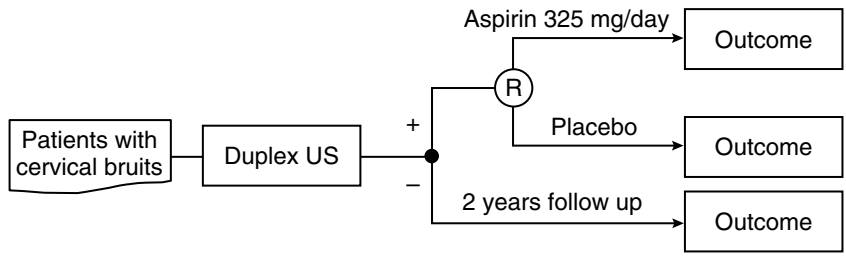


Figure 4.3 Randomising abnormal test results.

Table 4.3 Analysis of an RCT, randomising only abnormal test results

(a) Natural prognostic value

	Poor outcome	Favourable outcome	
Stenosis ≥50%*	130 (71%)	54 (29%)	184
Stenosis <50%	255 (78%)	72 (22%)	327
Total	385	126	511

(b) Treatment effect in case of ≥ 50% stenosis

	Poor outcome	Favourable outcome	
Aspirin	138 (73%)	50 (27%)	188
Placebo	130 (71%)	54 (29%)	184
Total	268	104	372

*Random sample of patients with a stenosis ≥50%.

Alternative randomised designs

An alternative to the design in Figure 4.3 would be to move the point of randomisation back in time to the point where the test results are not yet known. This comes down to the randomisation of all patients to either disclosure or non-disclosure of the test results.

The latter design was used to evaluate Doppler ultrasonography of the umbilical artery in the management of women with intrauterine growth retardation (IUGR).[7] A total of 150 pregnant women with IUGR underwent Doppler ultrasonography and were subsequently randomised to disclosure or non-disclosure of the test results (Figure 4.4(a)). In the group in which the results of the test were revealed, women were hospitalised in case of abnormal flow and discharged with outpatient management in case of normal flow. In the non-disclosure group all patients received the conventional strategy for women with IUGR, of hospitalisation regardless of their test results. The trial compared perinatal outcome, neurological development and postnatal growth between the two strategies. The trial design, depicted in Figure 4.4(a), allows us to determine the natural prognostic value and the treatment effect in test abnormals. Unfortunately the authors did not report sufficient data to reconstruct the necessary 2×2 tables.

One could move the point of randomisation further back in time, to the decision whether or not to perform the test. A translation of this

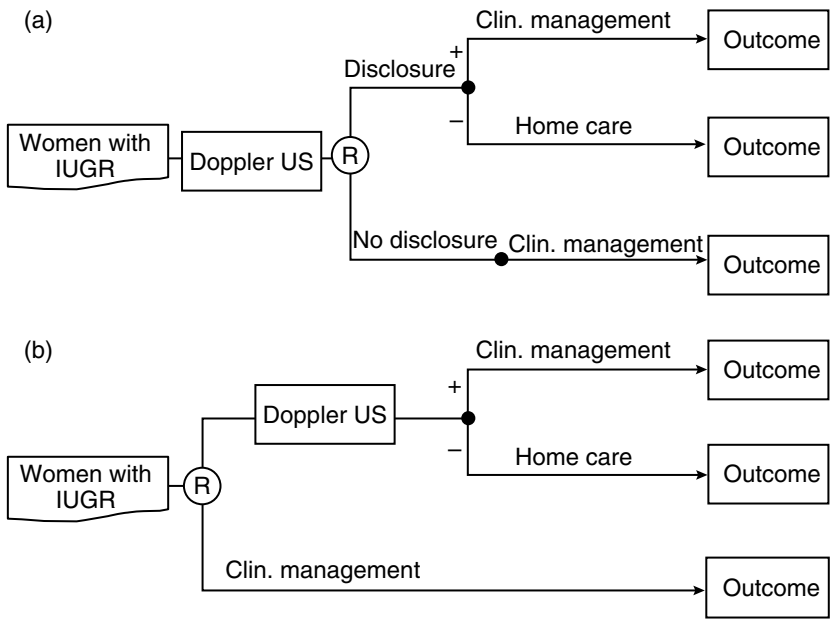


Figure 4.4 Alternative randomised designs.

comparison for the Doppler US in IUGR is the RCT in Figure 4.4(b). Women in whom IUGR has been diagnosed are randomly allocated to two strategies. The first consists of applying the test of interest, Doppler ultrasonography, in all women with IUGR. In case of abnormal flow a patient is hospitalised. In case of normal flow they are discharged with outpatient management. In the second strategy all women with IUGR are hospitalised. Subsequently, neonatal outcome is observed in each trial arm. A comparison of the outcomes in the two arms offers a measure of the effectiveness of using Doppler ultrasonography in making decisions on hospitalisation. Such a design evaluates both the test and the treatment effect; it is, however, not possible to distinguish the treatment effect from the prognostic value of the test.

Similar outcomes in both arms will be observed if there is no difference in outcome with either home care or clinical management in all patients satisfying the inclusion criteria for this trial. Differences in outcome are not necessarily attributed to the test. In case of a wrong choice of treatment, the outcome of the Doppler ultrasonography arm can turn out to be inferior to the conventional strategy, no matter how good or reliable the test actually is. This same line of reasoning can also be applied in case of a superior outcome in the Doppler arm. If there is a (sub)group of patients that is better off with home care, then the expected outcome in the Doppler ultrasonography group will always be superior, regardless of the intrinsic quality or accuracy of the test.

These examples demonstrate that it is not possible to make conclusions on the prognostic impact of the test itself using the design in figure 4(b), as long as it remains unclear to what degree results of such a trial depend on the new treatment, on accurate selection through the test, or both.

How to compare test strategies

In many clinical situations there are multiple tests available to examine the presence of the target condition. When one wishes to compare two competing tests the first three designs introduced earlier for the evaluation of a single test have to be adapted slightly.

To compare the prognostic value of two tests, a straightforward translation of Figure 4.1 is to perform both tests in all patients and to monitor the outcome of interest during a follow up period. Such a design is outlined in Figure 4.5(a). The data from such a study can be used to calculate and compare the prognostic value of each test, using conventional measures of diagnostic accuracy. One can also analyse the data by stratifying the results according to the possible test combinations. With two dichotomous tests this will result in a 4×2 table (Table 4.4). Note that each possible combination of results on test A and test B is treated as a separate test result category, analogously to a single test with four possible

result categories. Subsequently, the predictive value or the likelihood ratio of each result category can be calculated as a measure of prognostic value.[8]

To examine both tests in the context of subsequent clinical decision making, it is possible to randomise all patients between two treatment

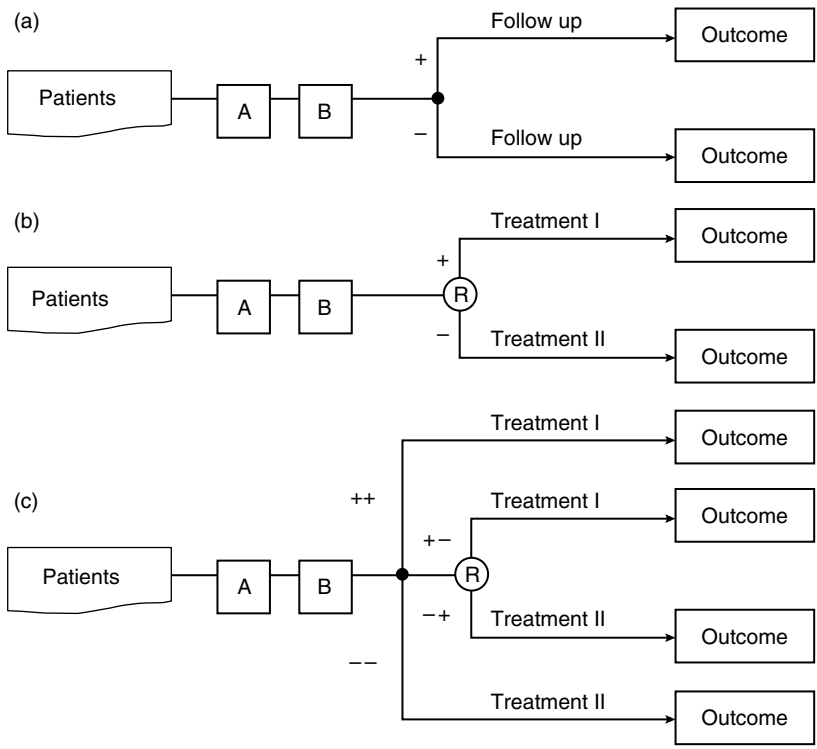


Figure 4.5 Designs to compare diagnostic strategies.

Table 4.4 A 4 × 2 table of the results of two dichotomous tests.

		Outcome	
A	B	+	−
+	+		
+	−		
−	+		
−	−		
Total			

strategies, similar to the design in Figure 4.2, regardless of their test results. Figure 4.5(b) shows an example of such a design: both tests are performed and all patients are randomly allocated to one of the two treatment options. This allows one to explore the prognostic value of both tests in each treatment arm. In addition, the data from such a trial can be used to find the most effective treatment for all patients included in the trial. If statistical power allows it, subgroup analysis of the treatment effect in the four possible test result categories offers the possibility of identifying the most effective treatment option for patients in the respective categories.

Although the previous design allows for many different explorations, only some are relevant from a clinical perspective. When two tests are compared, one of them is often already used in clinical practice and decisions on subsequent management are made based on this test. Let us assume that, in clinical practice, test positive patients are treated and test negative patients are not. If future decisions are to be made under the guidance of the new test, patients who are positive on the new test will be treated and those who are negative will not. This means that the only patients that will be managed differently are the ones who are test positive on the existing test but negative on the new one, and those who are test negative on the existing test but positive on the new one.

As patients with concordant test results (++ or −−) will receive the same management, it is unnecessary and in some circumstances even unethical to examine the treatment effect in these two subgroups. If a new test (B) is then examined with the goal of substituting the old, possibly more invasive, and/or costly test (A), the design in Figure 4.5(c), randomising only the discordant test results, is more efficient. Subsequently, the treatment effect and the predictive values of the discordant result categories (A+B− and A−B+) can be examined (see Table 4.5(a–d)).

By transposing these tables it is possible to examine the effect of a clinical pathway based on test A or test B for patients with discordant test results (Tables 4.5(e) to 4.5(f)). The difference in poor outcome rate between these two tables, after correcting for the frequency of discordant results, is equal to the absolute risk difference of a clinical pathway based on test A compared to a pathway based on test B. To calculate the relative risk or the total risk of each strategy separately it is necessary to have information on the clinical event rate in each concordant group.

An alternative design, using the random disclosure principle, is outlined in Figure 4.6(a). Both test A and test B are performed in all patients. Subsequently, patients are randomised between a clinical pathway based on test A without disclosing the results of test B, and a pathway based on test B with non-disclosure of the results of test A. The same measures and tables can be obtained from such a design as discussed for the design in Figure 4.5.

In some situations one might wish to let the point of randomisation coincide with that of the clinical decision to choose either test A or test B and

Table 4.5 Analysis of an RCT of two tests randomising only discordant results.

(a) Treatment effect A+B−

A+	B−	Poor outcome	Favourable outcome
Treatment I			
Treatment II			
Total			

(b) Treatment effect A−B+

A−	B+	Poor outcome	Favourable outcome
Treatment I			
Treatment II			
Total			

(c) Treatment

A	B	Poor outcome	Favourable outcome
+	−		
−	+		
Total			

(d) No treatment

A	B	Poor outcome	Favourable outcome
+	−		
−	+		

(e) Strategy based on A

A	B	Poor outcome	Favourable outcome
+	−		
−	+		
Total			

(f) Strategy based on B

A	B	Poor outcome	Favourable outcome
+	−		
−	+		

act on the respective results. A recent trial used this design to study two different diagnostic approaches for the management of outpatients with dysphagia.[9] Patients with dysphagia are at risk for aspiration pneumonia. A modified barium swallow test (MBS) and flexible endoscopic evaluation of swallowing with sensory testing (FEESST) are supposed to distinguish patients who can benefit from behavioural and dietary management from those who will need a percutaneous endoscopic gastrostomy (PEG) tube. For the discussion we consider a simplified design as outlined in Figure 4.6(b). Outpatients presenting with dysphagia were randomly allocated to either a strategy using MBS or a strategy using FEESST to guide subsequent management. During 1 year of follow up the occurrence of pneumonia was recorded in both trial arms. There were six cases of pneumonia in the 50 (12%) patients allocated to the FEESST strategy and 14 in the 76 (18%) patients allocated to the MBS strategy. The absolute risk difference was not significantly different from zero (risk difference 6%; 95% CI −6% to 19%). As no patient received both tests it is not possible to distinguish the

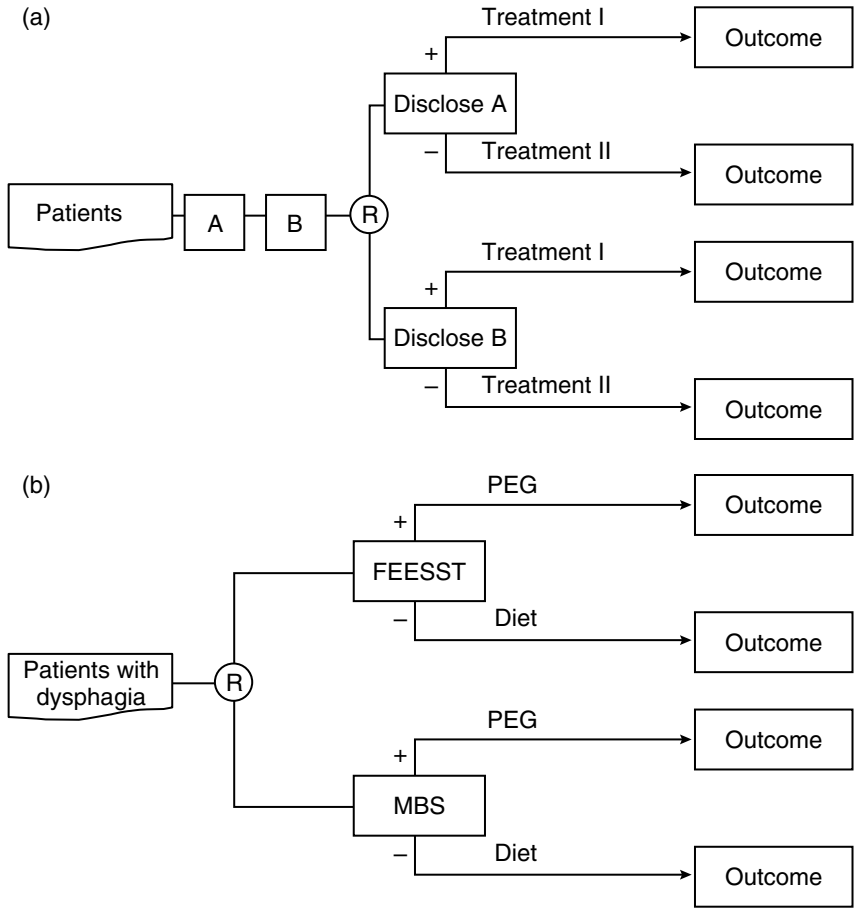


Figure 4.6 Alternative designs to compare diagnostic strategies.

treatment effect from the prognostic value of the tests, nor is it possible to compare the outcome in the subgroups with discordant test results.

Often a new test is introduced to complement rather than to replace existing tests. One example is where the new test is to be added to the diagnostic pathway before an older test as a triage instrument. Patients with a particular test result (say, negative) on the new test will not be subjected to the existing test. Alternatively, the new test is added after the existing test, making further refinement in diagnosis or treatment decisions possible. We refer to these two options as pre- and postaddition.

If a test is added at the end of a diagnostic work up to further classify disease (postaddition) all the designs presented in Figures 4.2 to 4.4 for the single test evaluation can be used to evaluate this new classification. For

example, to evaluate the prognostic impact of a genetic test for the classification of women with breast cancer in two different subgroups, one could use a design similar to the one in Figure 4.3. Women suspected of breast cancer are evaluated with the conventional diagnostic work up. Subsequently, only women with breast cancer are eligible for the trial. In all these women genetic tests are performed. Depending on the results, they are subsequently randomised between two types of treatment.

When the goal of a new test is to limit the number of people undergoing the classic diagnostic work up (triage or preaddition), the designs in Figures 4.5(b)–(c) and 4.6(a) can be used to evaluate the prognostic impact of such a strategy. Using the principle that only patients with test results that will actually account for the difference are randomised, one could also adapt the design of Figure 4.5(b), randomising only patients with the pair of discordant test results that will be treated differently if the new strategy is adopted. Another option is shown in Figure 4.7(a). As the difference between the two strategies comes from the group of patients who are not selected for the classic diagnostic work up, one can randomise only these patients to either the classic work up and treatment, or management based on the results of the new test.

Many studies to evaluate the preaddition of a test have randomised all patients between the two different diagnostic work ups.[10,11] One such study evaluated *Helicobacter pylori* serology as a way to reduce the number of patients subjected to endoscopy. Lassen et al.[10] performed the trial outlined in Figure 4.7(b). Patients presenting in primary care with dyspepsia were randomly assigned to either *H. pylori* and eradication therapy or prompt endoscopy. In case of a negative *H. pylori* test patients were still subjected to endoscopy. During a 1 year follow up the symptoms were recorded on a Likert scale.

Choice of design

Each of the designs discussed in Figures 4.1 to 4.7 has its own advantages and disadvantages. Depending on the clinical problem one wishes to answer, the type of information needed, and the costs of tests or follow up, one design can be preferred over another.

The outlines in Figures 4.2 to 4.4 can be used to evaluate a strategy with a new test compared to a classic strategy without such a test. In case of postaddition the classic strategy will consist of the classic diagnostic work up and treatment. In case of a substitution problem, any of the trial designs outlined in Figures 4.5(b) to 4.6(b) can provide an answer. The designs outlined in Figures 4.5(b), 4.6(a), 4.7(a), and 4.7(b) can provide an answer in case of a preaddition problem.

Table 4.6 gives an overview of the information that can be deduced from the different designs. The designs in Figures 4.2 and 4.5(b), testing all

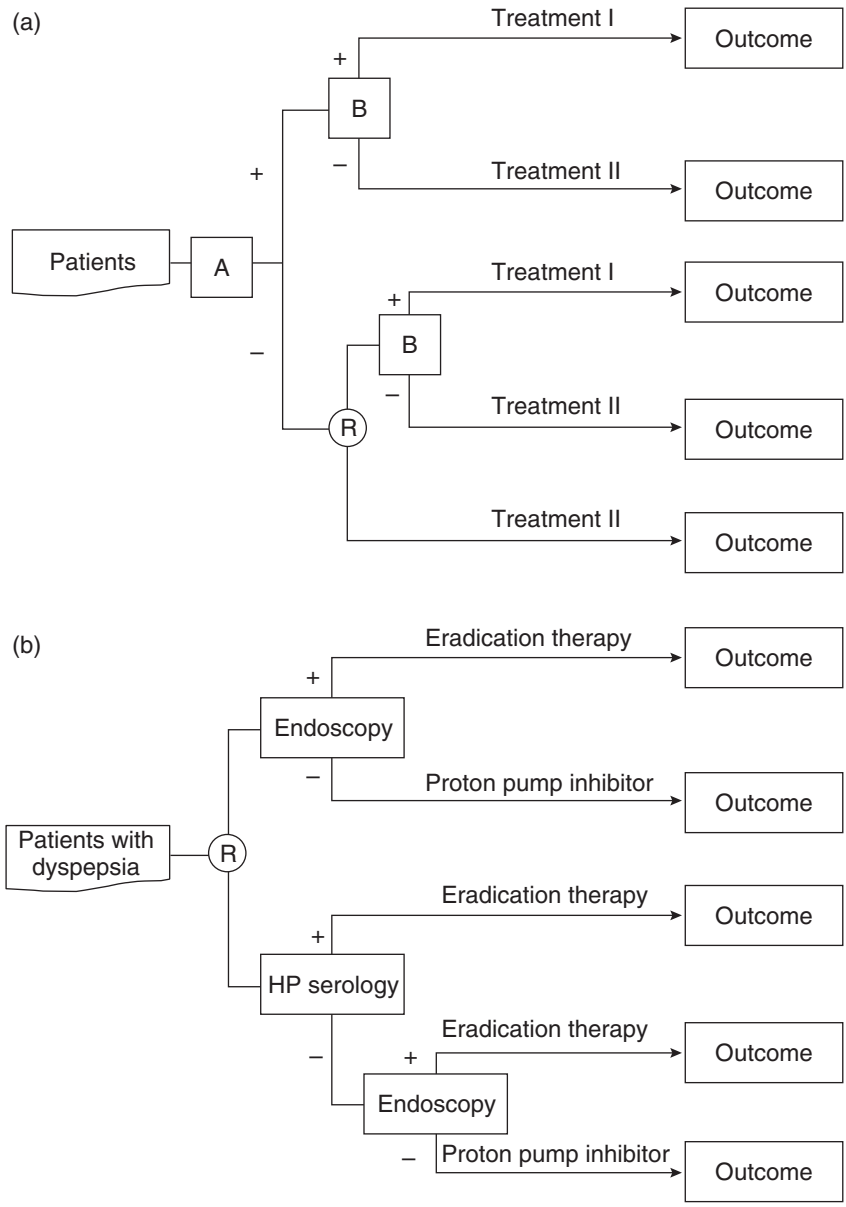


Figure 4.7 Designs to evaluate preaddition.

patients and randomising all between two treatment strategies, provide the most information. In addition to data on the effects of the two evaluated strategies, they provide information on the treatment effect and prognostic value of all possible test result categories. Yet these designs are not always

Table 4.6 Possible analyses of each randomised design.

Figure	1, 5(a)	2	3, 4(a)	5(b)	5(c) 6(a)	4(b) 6(b) 7(a) 7(b)
Natural prognostic value	×	×	×	×		
Prognostic value under intervention		×		×		
Treatment effect test abnormals		×	×	×		
Treatment effect test normals		×		×		
Treatment effect discordant tests				×	×	
Strategy effect		×	×	×	×	×

ethical, as there is often evidence that one treatment is better for some of the test result categories. In that case a better alternative are the designs outlined in Figures 4.3, 4.5(c) and 4.7(a), in which only the patients for whom there is uncertainty in the subsequent management are randomised.

The designs in Figures 4.4(b), 4.6(b) and 4.7(b) have frequently been used in the medical literature, probably because of their pragmatic attractiveness. In these designs the point of randomisation coincides with the decision to perform either test A or test B. From a cost perspective these designs can be more be economical than the other designs, in case of an expensive test, as on average fewer patients are tested than with the other designs. If follow up is expensive, designs randomising only patients with the test category of interest (Figures 4.3, 4.5(c) and 4.7) are more efficient, as fewer patients will be needed to achieve the same amount of statistical precision.[12] However, the latter designs are not feasible when tests that influence each other's performance are being compared. For example, it is not possible to compare two surgical diagnostic procedures, mediastinoscopy and anterior mediastinotomy, for the detection of mediastinal lymphomas by performing them both in all patients as suspected lymph nodes are removed.[13]

Practical issues

We have discussed the pros and cons of different designs to evaluate the prognostic impact of a single test or to compare different test strategies. In the design of a trial there are several other issues that should be considered in advance. In all of the examples we have presented here there was a prespecified link between test results and management decisions. Test positive patients were to receive one treatment, test negative another. If such a link is absent, and clinicians are free to select therapy for each test result, it will remain unclear to what extent poor trial results reflect deficiencies of the test itself, ineffective treatment options or, alternatively, incorrect management decisions. Detailed information on the treatment protocol is also necessary for others to implement the possible findings of the study.

A clear specification of the treatment options and their relation with the different test results is an absolute necessity for any diagnostic study.[12]

As for each randomised controlled trial, methods to preserve allocation concealment and blinding deserve special attention. It has been shown empirically that inadequate concealment of allocation, as well as inadequate blinding, can lead to exaggerated estimates of a strategy's effectiveness.[14] One way to guard adequate allocation concealment is a central randomisation procedure. In some situations the use of sealed opaque envelopes, with monitoring of the concealment process, may be more feasible.[15] Blinding of the outcome measurement for the randomisation outcome is of greater importance for some outcomes than for others, but can be implemented with the same methods as developed for therapeutic trials. Blinding of the clinician or patient to the allocation is more difficult. When two different strategies are randomised (Figure 4.6(b)) one can imagine that the knowledge of the type of test influences subsequent management decisions, despite a prespecified link. For example, an obstetrician might be more reassured with the results of magnetic resonance pelvimetry in a breech presentation at term, than with manual pelvimetry, which will influence subsequent decisions to perform an emergency section.[16] One could choose a design that randomises test results to overcome this problem. Alternatively, one could try to mask the clinician by only presenting standardised test results, without any reference to the type of test.

The *a priori* calculation of the necessary sample size for a randomised diagnostic study is not straightforward. When discussing Figure 4.5(c) we showed that the expected difference in outcome between the two test strategies resulted from the expected difference in the category with discordant test results only. In trials in which patients are randomised to one of two test strategies (Figure 4.6(b)) a large group of participants will also not contribute to the final difference. Let us explain this with another randomised diagnostic trial from the literature, in which ultrasonography was compared with clinical assessment for the diagnosis of appendicitis.[17] The authors report a power of 80% to detect a reduction in the non-therapeutic operation rate from 11% to 2%, by randomising 302 patients. What are the nominator and denominator of these estimated rates?

Figure 4.8 shows the two trial arms. A large group of patients with abnormal results in the ultrasound group, indicating operation, would also have been detected at clinical examination. The same argument stands for a subgroup of patients with a normal ultrasound. The sum of these two groups forms the total with concordant test results. As patients with concordant test results will receive the same management, their event rates will be identical except for chance differences. The rate of 11% results from $(x+n+o)/151$. The rate of 2% results from $(y+n+o)/151$. The rate difference, 9%, results solely from the events in the discordant group. By

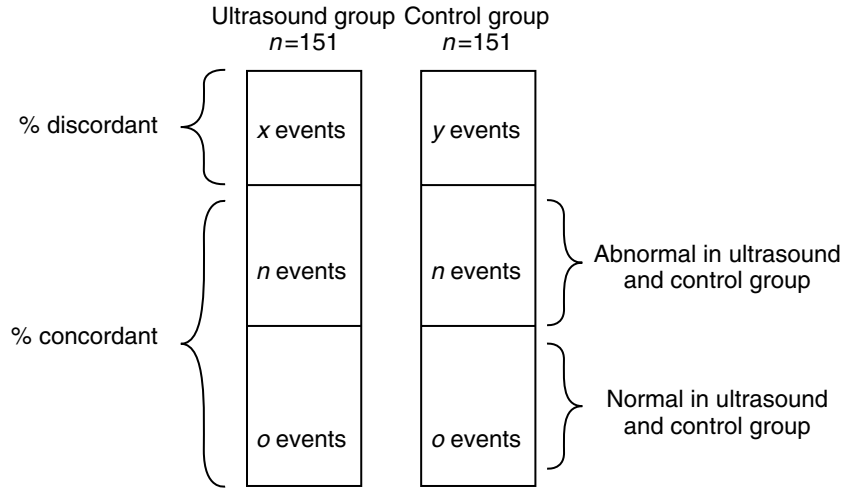


Figure 4.8 Sample size calculation.

assuming a concordance rate of ultrasonography with clinical assessment of 80%, one can calculate the postulated rate difference in this discordant group: 9%/20% is 45%. This could result from a rate of non-therapeutic operations of 55% in patients with a positive clinical assessment and otherwise negative ultrasound, and a rate of 10% in patients with a positive ultrasound and otherwise negative clinical examination. (This implies that the event rate is 0% in the concordant group, which is not very likely, as the authors discuss in their introduction that 15–30% of all operations are non-therapeutic.) With some extra calculations we can show that the difference assumed by the authors implies a discordance rate of at least 80%. It would be very strange to expect such a high discordance rate in advance. This example shows that it is important to incorporate the discordance rate in sample size calculations of randomised trials of diagnostic tests.

Conclusions

In this chapter we discuss the evaluation of the prognostic impact of tests. From a patient perspective one could argue that it is not so much the correspondence with the "truth" that should be the focus of a diagnostic test evaluation, but the likelihood that such a test detects events of clinical interest, and the possibilities that exist to let test results guide subsequent clinical decision making to reduce the likelihood of such events occurring. The latter can be evaluated by evaluating a test–treatment combination in a clinical trial, for which several possible designs are discussed.

The examples of published randomised diagnostic trials in this chapter show that it is feasible to perform such a thorough evaluation of a

diagnostic test. Several additional examples can be found in the literature, such as trials of mediastinoscopy, cardiotocography and MRI,[11,18,19] and of a number of screening tests.[20–22] These date back as far as 1975.[23]

In most of these trials the point of randomisation coincided with the clinical decision as to whether or not to perform the tests. This makes it impossible to differentiate between the treatment effect and the prognostic value of the test. Power analyses of any diagnostic trial should incorporate an estimation of the discordance rate, as differences in outcome can only be expected for patients who have discordant test results. In this chapter we have shown that a design incorporating randomisation of discordant test results is more efficient, provides more information, and is less prone to bias. Most importantly, all of these designs require a prespecified test–treatment link. This is to allow for the application of the study results in other settings, and to guard the internal validity of the study.

References

1 Rothwell PM, Pendlebury ST, Wardlaw J, Warlow CP. Critical appraisal of the design and reporting of studies of imaging and measurement of carotid stenosis. *Stroke* 2000;**31**:1444–50.

2 Lewis RF, Abrahamowicz M, Cote R, Battista RN. Predictive power of duplex ultrasonography in asymptomatic carotid disease. *Ann Intern Med* 1997;**127**:13–20.

3 Adams HP Jr, Bendixen BH, Leira E, *et al*. Antithrombotic treatment of ischemic stroke among patients with occlusion or severe stenosis of the internal carotid artery: a report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology* 1999;**53**:122–5.

4 Anonymous. Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke: a randomized controlled trial. The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) Investigators [see comments]. *JAMA* 1998;**279**:1265–72.

5 Cote R, Battista RN, Abrahamowicz M, Langlois Y, Bourque F, Mackey A. Lack of effect of aspirin in asymptomatic patients with carotid bruits and substantial carotid narrowing. The Asymptomatic Cervical Bruit Study Group. *Ann Intern Med* 1995;**123**:649–55.

6 Kraemer HC. *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park: SAGE Publications, 1992:295.

7 Nienhuis SJ, Vles JS, Gerver WJ, Hoogland HJ. Doppler ultrasonography in suspected intrauterine growth retardation: a randomized clinical trial. *Ultrasound Obstet Gynecol* 1997;**9**:6–13.

8 Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991;**44**:763–70.

9 Aviv JE. Prospective, randomized outcome study of endoscopy versus modified barium swallow in patients with dysphagia. *Laryngoscope* 2000;**110**:563–74.

10 Lassen AT, Pedersen FM, Bytzer P, de Muckadell OBS. *Helicobacter pylori* test-and-eradicate versus prompt endoscopy for management of dyspeptic patients: a randomised trial. *Lancet* 2000;**356**:455–60.

11 Anonymous. Investigation for mediastinal disease in patients with apparently operable lung cancer. Canadian Lung Oncology Group. *Ann Thorac Surg* 1995;**60**:1382–9.

12 Bossuyt P, Lijmer J, Mol B. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;**356**:1844–7.

13 Elia S, Cecere C, Giampaglia F, Ferrante G. Mediastinoscopy vs. anterior mediastinotomy in the diagnosis of mediastinal lymphoma: a randomized trial. *Eur J Cardiothorac Surg* 1992;**6**:361–5.

14 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.
15 Swingler GH, Zwarenstein M. An effectiveness trial of a diagnostic test in a busy outpatients department in a developing country: issues around allocation concealment and envelope randomization. *J Clin Epidemiol* 2000;**53**:702–6.
16 van der Post JA, Maathuis JB. Magnetic-resonance pelvimetry in breech presentation [letter; comment]. *Lancet* 1998;**351**:913.
17 Douglas C, Macpherson N, Davidson P, Gani J. Randomised controlled trial of ultrasonography in diagnosis of acute appendicitis, incorporating the Alvarado score. *BMJ* 2000;**321**:1–7.
18 Strachan BK, van Wijngaarden WJ, Sahota D, Chang A, James DK. Cardiotocography only versus cardiotocography plus PR-interval analysis in intrapartum surveillance: a randomised, multicentre trial. FECG Study Group. *Lancet* 2000;**355**:456–9.
19 Dixon AK, Wheeler TK, Lomas DJ, Mackenzie R. Computed tomography or magnetic resonance imaging for axillary symptoms following treatment of breast carcinoma? A randomized trial. *Clin Radiol* 1993;**48**:371–6.
20 Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet* 1996;**348**:1467–71.
21 Anonymous. Controlled trial of universal neonatal screening for early identification of permanent childhood hearing impairment. Wessex Universal Neonatal Hearing Screening Trial Group. *Lancet* 1998;**352**:1957–64.
22 Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-Year Results of a Randomized Trial in Women Aged 50–59 Years. *J Natl Cancer Inst* 2000;**92**:1490–9.
23 Morris DW, Levine GM, Soloway RD, Miller WT, Marin GA. Prospective, randomized study of diagnosis and outcome in acute upper-gastrointestinal bleeding: endoscopy versus conventional radiography. *Am J Dig Dis* 1975;**20**:1103–9.

5 The diagnostic before–after study to assess clinical impact

J ANDRÉ KNOTTNERUS, GEERT-JAN DINANT, ONNO P VAN SCHAYCK

Summary box

- The before–after design is more appropriate for evaluating the clinical impact of additional testing than to compare the impact of different diagnostic options.
- Demonstrating an effect of diagnostic testing on the patient's health outcome is more difficult than showing a change in the doctor's assessment and management plan.
- Whether and what specific blinding procedures have to be applied depends on the study objective.
- To optimise the assessment of the independent effect of the test information, performance of the test or disclosure of the test result can be randomised. This would change the before–after design in a randomised trial.
- The therapeutic consequences of the various test results can be standardised in the research protocol, provided that such therapy options are clinically rational and have a well documented evidence base. The study will then evaluate the impact of the test result connected with a defined therapeutic consequence, rather than the test result per se.
- If evaluating the doctor's assessment is the primary study objective, the assessment should preferably take place immediately after

disclosure of the test result, with a minimal risk of interfering factors influencing the doctor's judgement.
- Because a rather long follow up is mostly needed to estimate the impact of testing on the clinical course, the risk of interfering influences is substantial.
- Given that before–after studies can be carried out relatively fast, largely embedded in daily care, while RCTs are more complex or expensive, a well designed before–after study may sometimes be used to explore whether and how a diagnostic RCT should be performed. If an RCT is impossible or infeasible, or ethically unacceptable, a well designed before–after study can be the most suitable alternative.

Introduction

Apart from facilitating an accurate diagnosis, a test is aimed at causing change: starting from a baseline situation, applying the test and interpreting its outcome should result in a new situation. In fact, the most important justification for diagnostic testing is that it is expected to make a difference, by influencing clinical management and ultimately benefiting the patient's wellbeing. Accordingly, performing a test can be seen as an intervention that should be effective in bringing about a clinically relevant change.

In studying the clinical effect of a test result, the randomised controlled trial (RCT) is the strongest methodological design option,[1] and is dealt with in Chapter 4. However, although it is the paradigm for effectiveness research, an RCT cannot always be achieved. This is, for example, the case if randomly withholding a test or test result from patients or doctors is considered medically or ethically unacceptable. Difficulties may also arise if the diagnostic test is integrated in the general skills of the clinician, so that performing it cannot be randomly switched on and off in his or her head, nor simply assigned to a different doctor. This is especially problematic if at the same time patients cannot be randomly assigned to a doctor. This situation may, for instance, occur in studying the impact of diagnostic reasoning skills in general practice. Also, when an RCT is complex and expensive, or will last too long to still be relevant when the results become available, one may wish to consider a more feasible alternative.

One alternative that may be considered is the diagnostic before–after study.[2] This approach seems attractive, as it fits naturally within the clinical process and is easier to perform than the randomised trial. Therefore, this chapter will discuss the potentials, limitations, and pitfalls of this design option.

The research question

Example

In a study to assess the diagnostic impact of erythrocyte sedimentation rate (ESR) in general practice, 305 consecutive patients with aspecific symptoms for whom general practitioners (GPs) considered ESR testing necessary were included.[3] Before testing, the GPs were asked to specify the most likely diagnosis in each patient, and to assess whether or not this diagnosis was severe in the sense of malignant or inflammatory disease for which further management would be urgently needed. Subsequently, the ESR was independently performed and the result was made available to the GPs, who then again specified their (revised) diagnostic assessment. After 3 months, based on all the available medical information, a clinical assessment was carried out for each patient by an independent clinician not knowing about the pre- and post-test assessments for each patient, in order to establish a final diagnosis (reference standard).[4]

In Figure 5.1 the percentage of patients most likely having severe pathology according to the GP is presented before and after disclosure of the ESR result. There seems to be no relevant pre–post-test change. However, looking at Table 5.1, it is clear that there was a change in 32 patients: 17 from severe pathology to "other", and 15 from "other" to severe pathology.

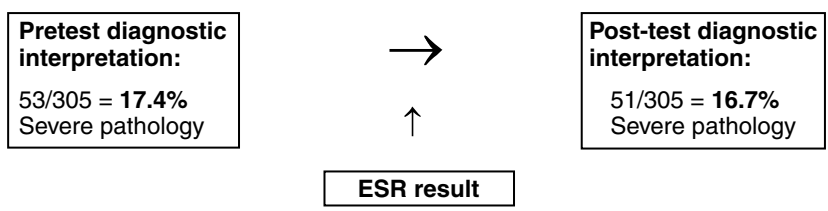


Figure 5.1 Pre- and post-test diagnostic assessment in studying the impact of ESR.

Table 5.1 Relation between pre- and post-test diagnostic assessments in studying the impact of ESR.

	Post-test interpretation		
Pretest interpretation	Severe pathology	Other	Total
Severe pathology	36	17	53
Other	15	237	252
Total	51	254	305

Whether these changes had indeed resulted in a more accurate diagnostic assessment could be determined after correlating the GPs' pre- and post-test findings with the reference standard procedure. It appeared that the pretest accuracy of the GPs' assessment was 69% (that is, 69% of cases were correctly classified), whereas the post-test accuracy was 76%, implying an increase of 7%.

Furthermore, of the 32 patients with a diagnostic classification changed by the GP, nine with a positive (severe) post-test diagnosis proved to be "false positives", and two with a negative (other) post-test diagnosis were "false negatives".

The test characteristics of the ESR (cut-off value ≥ 27 mm/1h) could be also determined in relation to the reference diagnosis, yielding a sensitivity of 53%, a specificity of 94%, a positive predictive value of 46%, and a negative predictive value of 91%.

The general model

The basic question in the diagnostic before–after study is whether applying a certain diagnostic test favourably influences the doctor's (a) diagnostic or (b) prognostic assessment of a presented clinical problem; (c) the further management; and, ultimately, (d) the patient's health. It essentially comprises the baseline (pretest) situation, a determinant (the test), and the outcome (post-test situation) (Figure 5.2).

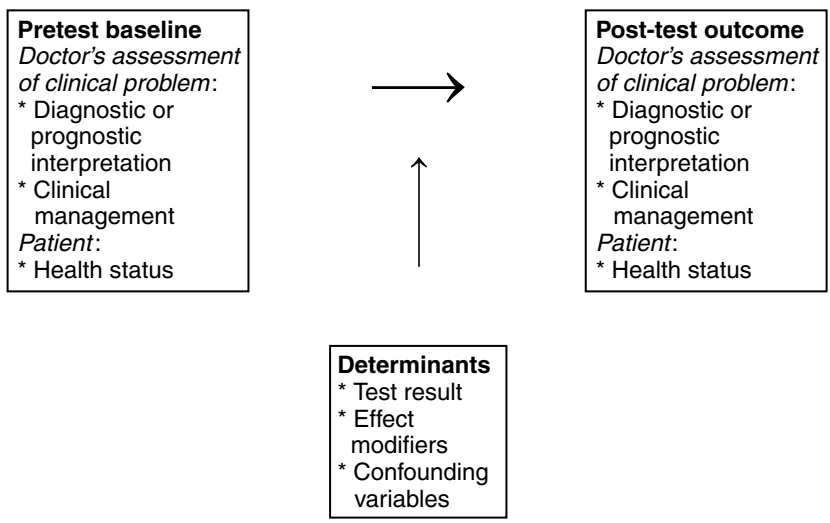


Figure 5.2 General representation of the research question in a diagnostic before–after study.

The point of departure can be characterised by a clinical problem, with the doctor's assessment regarding the possible diagnosis, prognosis, or the preferred management option, and the patient's health status at baseline, without knowing the information from the test to be evaluated. The patient's health status at baseline is important, not only as a starting point for possible outcome assessment but also as a reference for generalising the study results to comparable patient groups.

The determinant of primary interest is performing the diagnostic test and disclosure of its result, which is in fact the intended intervention. Furthermore, because diagnostic classification is essentially involved with distinguishing clinically relevant subgroups, it is often useful to consider the influence of effect modifying variables, such as the doctor's skills and experience, the patient's age and gender, and pre-existing comorbidity. In addition, the effect of possible confounding variables should be taken into account. For example, extraneous factors such as reading publications or attending professional meetings may affect the clinician's assessment. But also the time needed to do the test and obtain its result may be important, as it may be used to think and study on the clinical problem, and this will independently influence the assessment. Moreover, the patient's health status may have changed as a result of the clinical course of the illness, by interfering comorbidity and related interventions, by environmental factors, or by visiting other therapists. The patient's symptom perception may have been influenced by information from family, friends, or the media, or by consulting the internet. Also, the patient may claim to have benefited from a diagnostic intervention because he does not wish to disappoint the doctor.

The key challenge for the investigator is now to evaluate the extent to which applying the diagnostic test has independently changed the doctor's diagnostic or prognostic assessment of the presented clinical problem, the preferred management option, or the patient's health status. The latter will generally be influenced indirectly, via clinical management, but can sometimes also be directly affected, for example because the patient feels himself being taken more seriously by the testing per se. Moreover, patient self testing, which is nowadays becoming more common, can influence patient self management.

At this point, two important limitations of the before–after design must be emphasised. First, the design is more appropriate to evaluate the impact of "add on" technologies[2] (that is, the effect of additional testing) than to compare the impact of different diagnostic technologies or strategies. For the latter purpose one could, in principle, apply both studied technologies, for example colonoscopy and double contrast barium enema, in randomised order, to all included patients, and then compare the impact of disclosing the test results, again in random order, on the clinicians' assessment. Another example would be to subject patients to both CT and

NMR head scanning to study their influence on clinicians' management plans in those with suspected intracranial pathology. However, such comparisons are unrealistic, as the two tests would never be applied simultaneously in practice. Moreover, such studies are very burdensome for patients, not to say ethically unacceptable, and would make it virtually impossible to study the complication rate of each procedure separately.[5] When the various options are mutually exclusive, for example when comparing diagnostic laparotomy with endoscopy in assessing intra-abdominal pathology as to their adverse effects, a before–after design is clearly inappropriate. In such situations, a randomised controlled trial is by far the preferred option. Only when the compared tests can be easily carried out together without any problem for the patient, can they be applied simultaneously. This can be done, for instance, when comparing the impact of different blood tests using the same blood sample. However, when the disclosure of the results of the compared tests to the clinicians is then randomised, which would be a good idea, we are in fact in the RCT option.

Second, demonstrating an effect of diagnostic testing on the patient's health outcome is much more difficult than showing a change in the doctor's assessment and management plan. In fact, this is often impossible, as it usually takes quite some time to observe a health effect that might be ascribed to performance of the test. Controlling for the influence of the many possible confounders over time generally requires a concurrent control group of similar patients not receiving the test. However, a diagnostic before–after study could be convincing in case of: (1) studying a clinical problem with a highly predictable outcome in the absence of testing (such as unavoidable death in the case of imminent rupture of an aneurysm of the aorta), (2) while adding specific diagnostic information (an appropriate imaging technique) leading to a specific therapeutic decision (whether and how to operate) (3), which is aimed at a clearly defined short term effect (prevention of a rupture and survival), possibly followed by less specific long term effects (rehabilitation). However, such opportunities are extraordinary. Besides, although on the one hand some clinicians would consider such clinical situations to be self evident and not needing evaluation by research, others may still see room for dispute as to what extent clinical events are predictable or unavoidable.

Working out the study

Pretest baseline

The study protocol follows the elements of the research question.

At baseline, the clinical problem and the study question are defined. The clinical problem could be aspecific symptoms as presented in primary care, with the question being whether the ESR would contribute to the doctor's

diagnostic assessment,[3,4] or sciatica, in order to study whether radiography would affect therapeutic decision making.

The health status of each patient to be included is systematically documented, using standardised measurement instruments for the presented symptoms, patient history, physical examination, and further relevant clinical data.

Overseeing all available patient data, the doctor makes a first clinical assessment of the probability of certain diagnoses or diagnostic categories. In primary care, for example, the probability of a severe organic malignant or inflammatory, disorder can be assessed. This can be done for one specified diagnostic category, for a list of specified diagnoses, or in an open approach, just asking the differential diagnosis the doctor has in mind, with the estimated probability of each specific diagnostic hypothesis being considered.

Furthermore, the doctor is asked to describe the preferred diagnostic or therapeutic management plan, which can be done, again, according to a prepared list of items or as an open question.

At baseline, possibly relevant effect modifying variables should be considered. Often the general clinical experience of the clinicians and their specific expertise regarding the test under study are important. Furthermore, variables characterising important clinical subgroups can be assessed, and potential confounding factors have to be measured in order to be able to take these into account in the data analysis. Recording of covariables is sometimes difficult, for example for extraneous variables such as media exposure. Moreover, it cannot be excluded that important or even decisive factors are not identified or even foreseen.

Diagnostic testing

In performing the diagnostic test or procedure under study and revealing its outcome after the baseline assessment, different options can be considered depending on the specific study objective.

- If one wishes to assess the specific effect of the test information on the outcome, in addition to the pretest clinical information, the test result should be determined independently from the pretest information. This is especially relevant for test procedures with a subjective element in the interpretation of the result, such as patient interviews, auscultation, imaging, *x* ray films, and pathological specimens. Accordingly, those who interpret the test should not be aware of the pretest information. However, when patient history itself is the test to be evaluated, this will generally not be feasible.
- When it is important to limit possible confounding effects of a preoccupation of participating doctors with the expected relevance of a

certain test, the investigator may wish to obscure the performing of the evaluated test itself. This can theoretically be achieved by not telling the doctor in advance about what specifically is being evaluated, and by disclosing the test result while also providing information on a number of other items irrelevant for the studied comparison. However, such masking is difficult and often not feasible, or may be so much in conflict with clinical reality that the findings will not be relevant for practice. Finally, intentional obscuring of the specific research question will need the explicit approval of the medical ethics review board.

- If the investigator wishes to assess the diagnostic process as it is usually embedded in clinical practice, the interpretation of the test result can be carried out as usual without specific blinding procedures. However, particularly for tests with subjective elements in reading or interpretation of the results, this will imply that the independent contribution of the test cannot be reliably determined.
- To optimise the assessment of the independent effect of the test information, performance of the test or, even more precisely, disclosure of the test result, can be randomised so that half of the participants would and half would not get the result. In fact, this would change the before–after design into a randomised trial, which is discussed in Chapter 4.

Because a test is almost never perfect, applying it may produce misclassification. Even when most patients are more accurately classified if the clinician uses the test information, some patients with a correct pretest diagnosis may be incorrectly classified after testing, for example because of a false positive or false negative result. As this may have important negative consequences for those patients – for example when a false positive mammography would lead to unnecessary surgery – it is recommended to include evaluation of the actual disease status in the context of the before–after study. This also enables the investigator to determine test accuracy by relating the test result cross-sectionally to the disease status, established according to an acceptable reference standard.[5]

Post-test outcome

The measurement of the final post-test outcome after disclosure of the test result (diagnostic assessment, preferred management plan, and/or patient health status) should follow the same procedure as the baseline measurement. In doing so, both the doctor and the patient will generally remember the baseline status, implying that the post-test assessment of the doctor's differential diagnosis and management options cannot be blinded for the pretest assessment. This has probably been the case in the example of the diagnostic impact of the ESR measurement.

When one is evaluating the impact of adding the test information to already known clinical information at baseline in order to make a comprehensive assessment, lack of blinding is not always a principal problem. In fact, it is clinically natural and supported by Bayes's theorem to study the impact of the test result in the light of the prior probability. However, when clinicians are more or less "anchored" to their initial diagnostic assessment, they are biased in that they do not sufficiently respond to the test information in revising their diagnostic assessment. But even this can sometimes be acceptable for an investigator who deliberately aims to assess the impact of the test in clinical reality, where such anchoring is a common phenomenon,[6,7] and to study the influence of the test result in terms of confirming or refuting this anchoring. As doctors will vary in "anchoring", such evaluations need to include sufficient participating clinicians.

When the post-test outcome is patient status, objective assessment independent of both pretest status and the doctor's interpretations is a basic requirement.

If one evaluates a test which is already firmly accepted among the medical profession, the response of clinicians is in fact "programmed" by medical education, continuing medical education, or clinical guidelines. In such cases the investigator is studying the adherence to agreed guidelines rather than the independent clinical impact of the test result. On the other hand, the clinical impact of testing will not be easily detected if there is no clear relationship between revision of the diagnostic classification based on the test information, and the revision of the management plan.[2] This relation can indeed be unclear when doctors ignore the test information, or when the same test result may lead to a variety of management decisions, including doing nothing. The latter can, for example, be the case when laboratory tests are carried out in asymptomatic patients. As a remedy, the therapeutic consequences of the various test results can be standardised in the research protocol, provided that such therapy options are clinically rational and have a well documented evidence base. Accordingly, the study will then evaluate the impact of the test result connected with a defined therapeutic consequence, rather than the test result per se. On the other hand, when there is a lack of clarity beforehand as to the potential management consequences of performing a test, we should ask ourselves whether such testing should be evaluated or used at all.

The time factor

The interval between the pre- and post-test assessments should be carefully chosen. Generally, the interassessment period should be short if evaluating the doctor's assessment is the primary study objective: the assessment should preferably take place immediately after disclosure of the test result, with a minimal risk of interfering factors influencing the doctor's

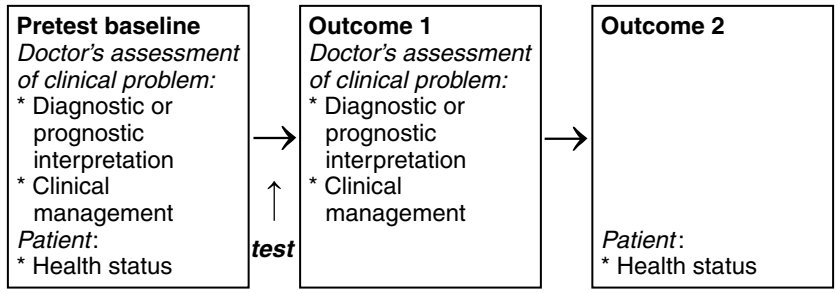


Figure 5.3 Separate post-test measurements of doctor's assessment (immediately) and patient health outcome (later).

judgement. Sometimes, however, this may take some days (bacterial culture) or longer (cervical smear). A rather long period until the final post-test assessment is mostly needed if estimating the impact of testing on the clinical course is the objective, although this longer period will be associated with an increased risk of interfering interventions and influences during follow up. A combined approach can be chosen, with a pretest measurement, a post-test measurement of the clinician's assessment, and a longer follow up period for measuring patient health outcome, respectively (Figure 5.3). In the analysis, then, the relation between the test's impact on the clinician's assessment and patient outcome could be studied if extraneous factors and changes in the clinical condition can be sufficiently controlled for. However, as outlined in the section on the research question, this is mostly impossible in the context of the before–after design. For the purpose of studying the test's impact on patient health, the randomised controlled trial is a more valid option.

Selection of the study subjects

Regarding the selection of the study subjects, similar methodological criteria to those discussed in Chapter 3 should be met: the study patient population should be representative for the "indicated", "candidate" or "intended" patient population, or target population, with a well defined clinical problem, clinically similar to the group of patients in whom the test would be applied in practice. Accordingly, the healthcare setting from where the patients come, the inclusion criteria, and the procedure for patient recruitment must be specified. Regarding the selection of participating doctors, the study objective is decisive. If the aim is to evaluate what the test adds to current practice, the pre- and post-test assessments should be made by clinicians representing usual clinical standards. However, if one wishes to ensure that the test's contribution is

analysed using a maximum of available expertise, top experts in the specific clinical field must be recruited.

Generally, in clinical studies a prospectively included consecutive series of patients with a clearly defined clinical presentation will be the most appropriate option with the lowest selection bias. If one were to retrospectively select patients who had already had the test in the past, one would generally not be able to be certain whether those patients really had a similar clinical problem, and whether all candidate patients in the source population would have non-selectively entered the study population. Apart from this, a valid before–after comparison of the doctors' assessments (with the doctors first not knowing and subsequently knowing the test result) is not possible afterwards, as a change in diagnostic assessment and the planning of management cannot be reliably reconstructed post hoc.

Sample size and analysis

Sample size requirements for the before–after study design need to be met according to general conventions. Point of departure can be the size of the before–after difference in estimated disease probability or other effectiveness parameters, for example, the decrease in the rate of (diagnostic) referrals, which would be sufficiently relevant to be detected. If the basic phenomenon to be studied is the clinical assessment of doctors, the latter are the units of analysis. When the consequences for the patients are considered the main outcome, their number is of specific interest.

The data analysis of the basic before–after comparison can follow the principles of the analysis of paired data. In view of the relevance of evaluating differences of test impact in various subgroups of patients, and given the observational nature of the before–after study, studying the effect of effect modifying variables and adjusting for confounding factors using multivariable analytical methods, may add to the value of the study. When the clinician and patient "levels" are to be considered simultaneously, multilevel analysis can be used.

As it is often difficult to reach sufficient statistical power in studies with doctors as the units of analysis, and because of the expected heterogeneity in observational clinical studies, before–after studies are more appropriate to confirm or exclude a substantial clinical impact than to find subtle differences.

Modified approaches

Given the potential sources of uncontrollable bias in all phases of the study, investigators may choose to use "paper" or videotaped patients or clinical vignettes, interactive computer simulated cases, or "standardised patients" especially trained to simulate a specific role consistently over time.

Standardised (simulated) patients can consult the doctor even without being recognised as "non-real".[8] Furthermore, the pre- and post-test assessments can also be done by an independent expert panel in order to ensure that the evaluation of the clinical impact is based on best available clinical knowledge. The limitations of such approaches are that they do not always sufficiently reflect clinical reality, are less suitable (vignettes) for an interactive diagnostic work up, cannot be used to evaluate more invasive diagnostics (standardised patients), and are not appropriate for additionally assessing diagnostic accuracy.

A before–after comparison in a group of doctors applying the test to an indicated patient population can be extended with a concurrent observational control group of doctors assessing indicated patients, without receiving the test information (quasi experimental comparison). However, given the substantial risk of clinical and prognostic incomparability of the participating doctors and patients in the parallel groups compared, and of possibly incorrectable extraneous influences, this will often not strengthen the design substantially. If a controlled design is considered, a randomised trial is to be preferred (Chapter 4).

Concluding remarks

As Guyatt et al.[2] have brilliantly pointed out, in considering a before–after design to study the clinical impact of diagnostic testing, two types of methodological problem must be acknowledged. First, we have to deal with problems for which, in principle, reasonable solutions can be found in order to optimise the study design. In this chapter, some of these "challenges" have been discussed. Examples are appropriate specifications of the clinical problem to be studied and the candidate patient population, and the concomitant documentation of test accuracy. Second, the before–after design has inherent limitations that cannot be avoided nor solved. If these are not acceptable, another design should be chosen. The most important of these limitations are: (1) the before–after design is especially appropriate for evaluating additional testing, rather than comparing two essentially different (mutually exclusive) diagnostic strategies; (2) the reported pretest management options may be different from the real strategy the clinicians would have followed if the test had not been available, or if they would not have known that there is a second (post-test) chance for assessment; (3) the pre- and post-test assessments by the same clinicians for the same patients are generally not independent; and (4) an unbiased evaluation of the impact of testing on the patients' health status can mostly not be achieved.

Acknowledging the large number of difficulties and pitfalls of the before–after design, as outlined in previous sections, we conclude that the design can have a place especially if the pre–post-test assessment interval

can be relatively short (evaluation of the test's impact on the doctor's assessment), and if the relation between the diagnostic assessment, the subsequent therapeutic decision making, and therapeutic effectiveness is well understood. If impact on patient outcome is studied, it is important that the clinical course of the studied problem in the absence of testing is well known and highly predictable.

Given the various limitations for studying the clinical impact of diagnostic tests, the randomised controlled trial design, if feasible, will in most cases be superior. However, given that before–after studies can be carried out relatively fast, largely embedded in daily care, whereas RCTs are more complex or expensive, a well designed before–after study may be useful to explore whether a diagnostic RCT could be worthwhile, or how it should be performed. In addition, if an RCT is impossible or infeasible, or ethically unacceptable, a before–after study may be the most suitable alternative. Other options, which could provide a more uniform clinical presentation and a better control of interfering variables, are before–after studies using written patient vignettes, interactive computer simulations, or standardised patients. The specific potentials and limitations (for example representing less clinical reality) of these alternative approaches will then have to be taken into account.

References

1 Alperovitch A. Controlled assessment of diagnostic techniques: methodological problems. *Effective Health Care* 1983;**1**:187–90.
2 Guyatt GH, Tugwell PX, Feeney DH, Drummond MF, Haynes RB. The role of before–after studies in the evaluation of therapeutic impact of diagnostic technology. *J Chronic Dis* 1986;**39**:295–304.
3 Dinant GJ, Knottnerus JA, van Wersch JWJ. Diagnostic impact of the erythrocyte sedimentation rate in general practice: a before–after analysis. *Fam Pract* 1992;**9**:28–31.
4 Dinant GJ. *Diagnostic value of the erythrocyte sedimentation rate in general practice*. Thesis. Maastricht: University of Maastricht, 1991.
5 Guyatt G, Drummond M. Guideline for the clinical and economic assessment of health technologies: the case of magnetic resonance. *Intl J Health Tech Assess Health Care* 1985;**1**: 551–66.
6 Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974; **185**:1124–31.
7 Elstein AS, Shulman LS, Sprafka SA. *Medical problem solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press, 1978.
8 Rethans JJ, Sturmans F, Drop R, van der Vleuten C. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *Br J Gen Pract* 1991; **41**:97–9.

6 Designing studies to ensure that estimates of test accuracy will travel

LES M IRWIG, PATRICK M BOSSUYT,
PAUL P GLASZIOU, CONSTANTINE GATSONIS,
JEROEN G LIJMER

Summary box

- There may be genuine differences between test accuracies in different settings, such as primary care or hospital, in different types of hospital, or between countries.
- Deciding whether estimates of test accuracy are transferable to other settings depends on an understanding of the possible reasons for variability in test discrimination and calibration across settings.
- The transferability of measures of test performance from one setting to another depends on which indicator of test performance is to be used.
- Real variation in the performance of diagnostic tests (such as different test types, or a different spectrum of disease) needs to be distinguished from artefactual variation resulting from study design features. These features include the target condition and reference standard used, the population and the clinical question studied, the evaluated comparison, and the way the index test was performed, calibrated, and interpreted.
- In preparing studies on diagnostic accuracy, a key question is how to design studies that carry more information about the transferability of results.

- In order to ensure that estimates of diagnostic accuracy will travel, before starting to design a study the following questions must be answered:
  - How are the target condition and reference standard defined?
  - Is the objective to estimate global test performance or to estimate probability of disease in individuals?
  - What is the population and clinical problem?
  - Is the test being considered as a replacement or incremental test?
  - To what extent do you want to study the reasons for variability of the results within your population?
  - To what extent do you want to study the transferability of the results to other settings?
- Designing studies with heterogeneous study populations allows exploration of the transferability of diagnostic performance in different settings. This will require larger studies than have generally been carried out in the past for diagnostic tests.

Introduction

Measures of test accuracy are often thought of as fixed characteristics that can be determined by research and then applied in practice. Yet even when tests are evaluated in a study with adequate quality – including features such as consecutive patients, a good reference standard, and independent, blinded assessments of tests and the reference standard[1] – diagnostic test performance in one setting may vary from the results reported elsewhere. This has been explored extensively for coronary artery disease,[2–5] but has also been shown for a variety of other conditions.[6–8] This variability is not only due to chance. There may be genuine differences between test accuracy in different settings, such as primary care or hospital, different types of hospital, or the same type of hospital in different countries. As a consequence, the findings from a study may not be applicable to the specific decision problem for which the reader has turned to the literature.

We suggest that deciding whether the estimates of test accuracy from studies are transferable to other settings depends on an understanding of the possible reasons for variability in test discrimination and calibration across settings. Variability may be due to artefactual differences (for example different design features of studies in different settings) or true differences (such as different test types, or a different spectrum of disease). To decide on the transferability of test results, we are concerned with true differences, after artefactual differences have been addressed.[9–11]

This chapter is divided into two main sections. The first is concerned with the reasons for true variability in accuracy; it explores conceptual

underpinnings. The second section is a pragmatic guide for those interpreting and designing studies of diagnostic tests. It is based on the view that value can be added to studies of diagnostic tests by exploring the extent to which we can characterise the reasons for variability in diagnostic performance between patients in different settings, and examining how much variability remains unexplained.

1. Reasons for true variability in test accuracy: conceptual underpinnings

Measures of diagnostic test performance: discrimination and calibration

There are many measures of test accuracy. Broadly speaking, we can think of them as falling into one of the following categories.

1 Global measures of test accuracy assess only discriminatory power. These measures assess the ability of the test to discriminate between diseased and non-diseased individuals. Common examples are the area under the receiver operating characteristic curve, and the odds ratio, sometimes also referred to as the diagnostic odds ratio. They may be sufficient for some broad health policy decisions, for example whether a new test is in general better than an existing test for that condition.

2 Measures of test performance to estimate the probability of disease in individuals require discrimination and calibration. These measures are used to estimate probabilities of the target condition in individuals who have a particular test result. An example is the predictive value: the proportion of people with a particular test result who have the disease of interest. To be useful for clinical practice, these estimates should be accompanied by other relevant information. For example, fracture rates in people with a particular result of a test for osteoporosis differ between people depending on their age, sex, and other characteristics. It is clumsy and difficult to estimate disease rates for all categories of patient who may have different prior probabilities. Therefore, the estimation is often done indirectly using Bayes' theorem, based on the patient-specific prior probability and some expression of the conditional distributions of test results: the distribution of test results in subjects with and without the target condition. Examples are the sensitivity and specificity of the test, and likelihood ratios for test results. These measures of test performance require more than the *discrimination* assessed by the global measures. They require tests to be *calibrated*. As an example of the difference between discrimination and calibration, consider two tests with identical odds ratios (and ROC curves) which therefore have the same discriminatory power. However, one test may operate at a threshold that gives a sensitivity of 90% and a

specificity of 60%, whereas the other operates at a threshold that gives a sensitivity of 60% and a specificity of 90%. Therefore, they differ in the way they are calibrated.

Features that facilitate transferability of test results

The transferability of measures of test performance from one setting to another depends on which indicator of test performance is to be used. The possible assumptions involved in transferability are illustrated in Figure 6.1. Table 6.1 indicates the relationship between these assumptions and the transferability of the different measures of test performance.

The main assumptions in transferring tests across settings are:

1 *The definition of disease is constant*. Many diseases have ambiguous definitions. For example, there is no single reference standard for heart failure, Alzheimer's disease or diabetes. Reference standards may differ because conceptual frameworks differ between investigators, or because it is difficult to apply the same framework in a standardised way.
2 *The same test is used*. Although based on the same principle, tests may differ – for example over time, or if made by different manufacturers.
3 *The thresholds between categories of test result (for example positive and negative) are constant*. This is possible with a well standardised test that can be calibrated across different settings. However, there may be no accepted means of calibration: for example different observers of imaging tests may have different thresholds for calling an image "positive". The effect of different cut points is classically studied by the use of an ROC curve. In some cases calibration may be improved by using category specific likelihood ratios, rather than a single cut point.
4 *The distribution of test results in the disease group is constant in shape and location*. This assumption is likely to be violated if the spectrum of disease changes: for example, a screening setting is likely to include earlier disease, for which test results will be closer to a non-diseased group (hence a lower sensitivity).
5 *The distribution of test results in the non-disease group is constant in shape and location*. This assumption is likely to be violated if the spectrum of non-disease changes: for example the secondary care setting involves additional causes of false positives due to comorbidity, not seen in primary care.
6 *The ratio of disease to non-disease (pretest probability) is constant*. If this were the case, we could use the post-test probability ("predictive" values) directly. However, this assumption is likely to be frequently violated: for example, the pretest probability is likely to be lowest in screening and greatest in referral settings. This likely non-constancy is the reason for using Bayes' theorem to "adjust" the post-test probability for the pretest probability of each different setting.

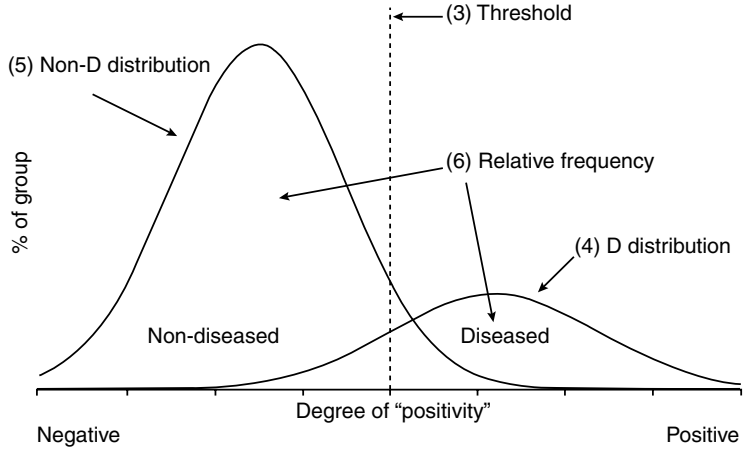


Figure 6.1 Distribution of test results in individuals with the disease of interest (D) and those without it (non-D). Numbers refer to assumptions for transferability of test results as explained in the text and Table 6.1.

Table 6.1 Assumptions for transferring different test performance characteristics. More important assumptions are marked **X** and those that are less crucial are marked *X*.

	Assumption*				
Measures of test discriminatory power	3	4	5	6	Comment
Odds ratio	**X**	**X**	**X**		Both of these measures are used for global assessment of discriminatory power and are transferable if the assumptions are met. Neither of them is concerned with calibration and therefore cannot be used for assessing the probability of disease in individuals
Area under ROC		**X**	**X**		
Measures of discriminatory power and calibration	3	4	5	6	
Predictive value	**X**	**X**	**X**	**X**	Directly estimates probability of disease in individuals
Sensitivity	**X**	**X**		*X*	
Specificity	**X**		**X**	*X*	These three measures can be used to estimate the probability of disease in individuals using Bayes' theorem
Likelihood ratios for a multicategory test	**X**	**X**	**X**		

*Assumptions are numbered as described in the text.

All the measures of test performance need the first two assumptions to be fulfilled. The extent to which the last four assumptions are sufficient is shown in Table 6.1, although they may not be necessary in every instance; occasionally the assumptions may be violated, but because of compensating differences transferability is still reasonable.

Lack of transferability and applicability of measures of test performance

We need first to distinguish artefactual variation from real variation in diagnostic performance. Artefactual variation arises when studies vary in the extent to which they incorporate study design features, such as whether consecutive patients were included, or whether the reference standard and the index test were read blind to each other. Once such artefactual sources of variation have been ruled out, we may explore the potential sources of true variation.[12] The issues to consider are similar to those for assessing interventions. For interventions, we consider patient, intervention, comparator, and outcome (PICO).[13,14] For tests the list is as follows, but with the target condition (equivalent to outcome in trials) shifted to the beginning of the list: (1) The target condition and reference standard used to assess it; (2) the population/clinical question; (3) the comparison; and (4) the index test. We now look at each of these in turn.

The target condition and the reference standard used to assess it

Test accuracy in any population will depend on how we define who has the target condition(s) that the test aims to detect. Clearly, the stage and spectrum of the target disease will influence the accuracy of the index test, as described later. However, even within a fixed spectrum and stage, there may be different definitions of who is "truly" diseased or not. Depending on the purpose of the study, the target conditions may be defined on grounds of clinical relevance, oriented to management decisions or prognosis, or defined on the grounds of pathological diagnosis. The definition of the target condition is therefore an active choice to be made by the investigator and its relevance interpreted by the reader of the study in the light of how they want to use the information. For example, should myocardial infarction include (a) "silent" myocardial infarction (with no chest pain)? (b) coronary thrombosis reversed by thrombolytic treatment, which then averts full infarction? This issue of the definition of the target condition and its method of ascertainment will clearly affect the apparent accuracy of the index test. For example, in parallel to considerations in clinical trials, the closer the reference standard is to a patient-relevant measure, the more this will help decisions about clinical applicability. Often reference standards that are considered objective and free of error are surrogates for (predictors of)

natural history, which could be measured directly. Consider the reference standard for a test for appendicitis. The "objective" reference standard for a new test of appendicitis is often considered to be histology (arrow 2 on Figure 6.2.) In fact, conceptually, follow up of natural history is a far more useful reference standard than histology (Figure 6.2, arrow 1). It is patient-relevant: those people who would have been found to have abnormal histology but whose condition resolves without operation can be considered false positives of the histological reference standard (Figure 6.2, arrow 3). In practice, the data for arrow 3 cannot be established, and we need to use a combined reference standard that we would consider as natural history when available and histology when not, rather than (as it is usually conceptualised) histology when available and natural history when not.

The usual presentation deals with a dichotomous definition of the target condition: it is either present or absent. In most cases the possibility of multiple conditions is more plausible. If these are known in advance, the polytomous nature can be taken into account.[15–17]

Misclassification of the reference standard will tend to result in underestimation of test accuracy if the errors in the reference standard and test are uncorrelated. The degree of underestimation is prevalence dependent in a non-linear way. Estimation of sensitivity is underestimated most when the prevalence of the target condition is low, whereas specificity is underestimated most when the prevalence of the target condition is high.[18,19] The odds ratio is underestimated most when prevalence is at either extreme. Therefore, error in the reference standard may cause apparent (rather than real) effect modification of test discrimination in subgroups in which the target condition has different prevalences.[20,21] This is shown in Table 6.2 where the same hypothetical test and reference standard are applied to a population in which disease prevalence is 50%

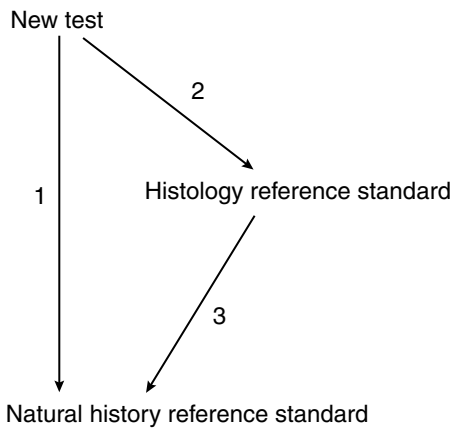


Figure 6.2 Choosing a relevant reference standard.

Table 6.2 Reference standard misclassification results in underestimation of test accuracy and apparent effect modification of different prevalences.

If reference standard has sensitivity = 0.9 and specificity = 0.8

	True disease				Reference standard		
Test	Present	Absent	Total	Test	Present	Absent	Total
Positive	80	30	110	Positive	78	32	110
Negative	20	70	90	Negative	32	58	90
Total	100	100	200	Total	110	90	200
Sensitivity = 0.80 Specificity = 0.70		OR = 9.3		Sensitivity = 0.71 Specificity = 0.64		OR = 4.4	

If reference standard has sensitivity = 0.9 and specificity = 0.8

	True disease				Reference standard		
Test	Present	Absent	Total	Test	Present	Absent	Total
Positive	80	300	380	Positive	132	248	380
Negative	20	700	720	Negative	158	562	720
Total	100	1000	1100	Total	290	810	1100
Sensitivity = 0.80 Specificity = 0.70		OR = 9.3		Sensitivity = 0.46 Specificity = 0.69		OR = 1.9	

(top half of table) and about 9% (bottom half). Sensitivity is reduced more in the population with 9% prevalence of disease, and specificity more in the population at 50% prevalence. The odds ratio is reduced most in the population at 9% prevalence. If errors in the reference standard are correlated with test errors, then the effect will be more difficult to predict. Correlated errors may result in overestimation of test accuracy.

The population and the clinical question

The population/clinical question are concerned not only with what disease is being tested for, but with what presentation of symptoms, signs and other information has prompted the use of the test. Test performance may vary in different populations and with minor changes in the clinical question. There are three critical concepts that help in understanding why this occurs. These are the spectrum of disease, the referral filter, and the incremental value of the test.

- *Spectrum of disease and non-disease*. Many diseases are not on/off states, but represent a spectrum ranging from mild to severe forms of disease.[22] Tumours, for example, start small, with a single cell, and then grow, leading eventually to symptoms. The ability of mammography, for example, to detect a breast tumour depends on its size. Therefore, test sensitivity will generally differ between asymptomatic and symptomatic

persons. If previous tests have been carried out the spectrum of disease in tested patients may be limited, with patients who have very severe forms of disease or those with very mild forms being eliminated from the population. For example, in patients with more severe urinary tract infection, as judged by the presence of more severe symptoms and signs, the sensitivity of dipstick tests was much higher than in those with minor symptoms and signs.[6]

Likewise, patients without the target condition are not a homogeneous group. Even in the absence of disease, variability in results is the norm rather than the exception. For many laboratory tests, normal values in women differ from those in men. Similarly, values in children differ from those in adults, and values in young adults sometimes differ from those in the elderly.

Commonly, the "non-diseased" group consists of several different conditions, for each of which the test specificity may vary. The overall specificity will depend on the "mix" of alternative diagnoses: the proportion of people in each of the categories that constitute the non-diseased; for example, prostate specific antigen may have a lower specificity in older people or those with prostatic symptoms, as it is elevated in men with benign prostatic hypertrophy.[23] In principle, patients without that target condition could represent a wide range of other conditions. However, the decision to use a test is usually made because of the presenting problem of the patient and the route by which they reached the examining clinician. Hence, the actual range of variability in patients without the target condition will depend on the mechanism by which patients have ended up in that particular situation. As an example, consider a group of ambulant outpatients presenting with symptoms of venous thromboembolism without having this disease compared to a group of inpatients suspected of venous thromboembolism but actually having a malignancy. The specificity of a D-dimer test in outpatients will be lower than that in inpatients.[24]

- *Referral filter*. The discriminatory power of tests often varies across settings because patients presenting with a clinical problem in one setting – for example primary care – are very different from those presenting to a secondary care facility with that clinical problem.[25,26] Patients who are referred to secondary care may be those with a more difficult diagnostic problem, in whom the usual tests have not resolved the uncertainty. These patients have been through a referral filter to get to the tertiary care centre.

This concept can best be considered using the hypothetical results of a diagnostic test evaluation in primary care (Table 6.3). Imagine that patients are referred from this population to a source of secondary care, and that all the test positive patients are referred, but only a random half of the test negative patients. As shown in Table 6.4, the overall test discrimination, as reflected in the odds ratio, has not changed. However, there appears to be a shift in threshold, with an increased sensitivity and a decreased specificity.

Of course, it is unlikely that test negatives would be referred randomly; rather, it may be on the grounds of other clinical information that the practitioner is particularly concerned about those test negatives. If the practitioner is correct in identifying patients about whom there is an increased risk of disease, the table could well turn out like Table 6.5.

In this case, because of the clinician's skill and the use of other information, not only does the test threshold appear to be shifted, but the overall test performance of the test in secondary care has been eroded, as shown by the reduced odds ratio. The more successfully the primary care practitioner detects cases that are test negative but which nevertheless need

Table 6.3 Accuracy of a test in primary care.

		Disease		
Test		Present	Absent	Total
Positive		60	40	100
Negative		40	60	100
Total		100	100	200
Sensitivity = 0.60 Specificity = 0.60	OR = 2.25			

Table 6.4 Test accuracy if a random sample of test negatives are referred for verification.

		Disease		
Test		Present	Absent	Total
Positive		60	40	100
Negative		20	30	50
Total		80	70	150
Sensitivity = 0.75 Specificity = 0.43	OR = 2.25			

Table 6.5 Diagnostic performances vary by setting because of selective patient referral.

		Disease		
Test		Present	Absent	Total
Positive		60	40	100
Negative		25	25	50
Total		85	65	150
Sensitivity = 0.71 Specificity = 0.38	OR = 1.5			

referral for management of the disease of interest, the more the performance of the test in secondary care is eroded.

- *To what prior tests is the incremental value of the new test being assessed*?

In many situations several tests are being used and the value of a particular test may depend on what tests have been done before,[27] or simple prior clinical information.[28,29] In Table 6.6 two tests are cross-classified within diseased and non-diseased people. The sensitivity and specificity of each test is 0.6, and they remain 0.6 if test B is used after test A, that is, the test performance characteristics of B remain unaltered in categories of patients who are A positive and those who are A negative.

However, if the tests are conditionally dependent or associated with each other within diseased and non-diseased groups, for example because they both measure a similar metabolite, then the overall test performance of B is eroded, as judged by the OR changing from 2.25 to 2.00 (Table 6.7 and Figure 6.3). In addition, there appears to be a threshold shift: the test is more sensitive but less specific in patients for whom A is positive than in those for whom A is negative. In other words, not only is the *discrimination* of the new test (B) less if done after the existing test (A), as judged by the odds ratio, but the *calibration* appears to differ depending on the result of the prior test. In fact, the threshold has not altered but there has been a

Table 6.6 Incremental value when tests A and B are conditionally independent.

	Have disease			No disease		
	A+	A−	Total	A+	A−	Total
B+	36	24	60	16	24	40
B−	24	16	40	24	36	60
Total	60	40	100	40	60	100

"Crude" sensitivity and specificity of both A and B = 0.6 and odds ratio = 2.25.
If A is + or −, SnB = 0.6, SpB = 0.6 and OR = 2.25.

Table 6.7 Incremental value when tests A and B are conditionally dependent.

	Have disease			No disease		
	A+	A−	Total	A+	A−	Total
B+	40	20	60	20	20	40
B−	20	20	40	20	40	60
Total	60	40	100	40	60	100

"Crude" sensitivity and specificity of both A and B = 0.6.
Odds ratio = 2.25.
If A +, SnB = 0.67, SpB = 0.50 and OR = 2.00.
If A −, SnB = 0.50, SpB = 0.67 and OR = 2.00.

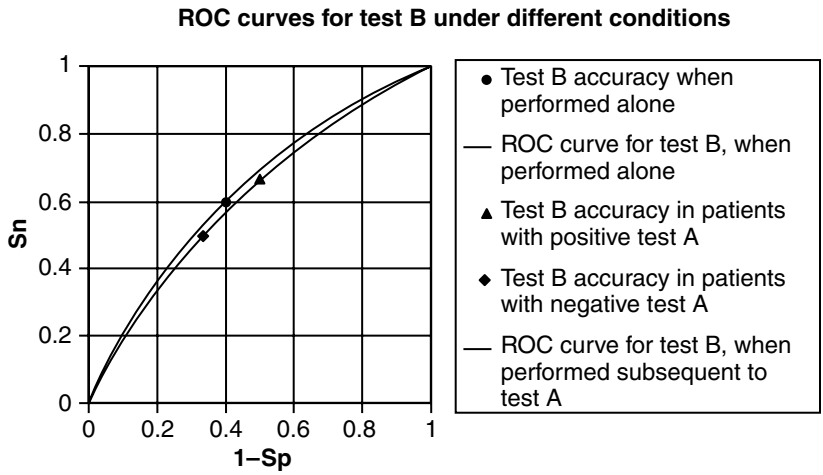


Figure 6.3 Test characteristics for test B alone and in those with positive and negative test A.

shift in the distribution of test results in diseased and non-diseased groups, conditional on the results of test A.

An example is provided by Mol and colleagues,[30] who evaluated the performance of serum hCG (human chorionic gonadotrophin) measurement in the diagnosis of women with suspected ectopic pregnancy. Several studies have reported an adequate sensitivity of this test.[30] However, the presence of an ectopic or intrauterine pregnancy can also be diagnosed with ultrasound. Mol et al. reported the sensitivity of hCG to be significantly different in patients with signs of an ectopic pregnancy (adnexal mass, or fluid in the pouch of Douglas) on ultrasound, compared to those without signs on ultrasound. As a consequence, an uncritical generalisation of the "unconditional" sensitivity will overestimate the diagnostic performance of this test if it is applied after an initial examination with ultrasound, as is the case in clinical practice.[30]

Categories of patients for whom new tests are most helpful are worth investigating. For example, whole-body positron emission tomography (PET) contributed most additional diagnostic information in the subgroup of patients in whom prior conventional diagnostic methods had been equivocal.[31]

The comparison: replacement or incremental test

Note that a new test may be evaluated as a *replacement* for the existing test, rather than being done after the existing test, in which case the *incremental* value is of interest. For assessment of replacement value, the cross-classification of the tests is not necessary to obtain unbiased estimates of how the diagnostic performance of the new test differs from that of the existing

one. However, information about how they are associated from a cross-classification will provide extra useful information and improve precision.[32]

Readers may have noticed that the issue of incremental value and the decreased test performance if tests are conditionally dependent is related to the prior issue of decreased test performance if the primary care clinician is acting as an effective referral filter. In our previous example, imagine that the test being evaluated is B. The clinician may be using test A to alter the mix of A+ and A−s that get through to secondary care, and the test performance of B reflects the way in which this mix has occurred.

The test

- *Discriminatory power*. Information about the test is relevant to both discriminative power and calibration. Discrimination may differ between tests that bear the same generic name but which, for example, are made by different manufacturers. Tests may be less discriminatory when produced in "kit" form than in initial laboratory testing.[33] When tests require interpretative skill, they are often first evaluated in near-optimal situations. Special attention is usually devoted to the unambiguous and reproducible interpretation of test results. This has implications for the interpretation and generalisability of the results. If the readers of images are less than optimal in your own clinical setting, test accuracy will be affected downward.[34–36]

The usual presentation deals with a two-way definition of test results, into positive and negative. In many cases, *multiple categories of test results* is more plausible. In addition, there may be a category of uninterpretable test results that needs to be considered. The polytomous nature of tests should be taken into account, for which several methods are available. Rather than a simple positive–negative dichotomy and the associated characteristics sensitivity and specificity, likelihood ratios for the multiple categories and ROC curves can be calculated (see Chapter 7). In all cases, a more general $n \times n$ table can be used to describe test characteristics, and several likelihoods can be calculated.[16]

- *Calibration*. If the purpose of the study is clinical decision making, in which information is being derived to estimate probablilities of disease, then a second major issue is the calibration of test results. A continuous test may have equivalent ROCs and diagnostic ORs in two different settings, but very different likelihood ratios (LRs). For example, machine calibration may be different in the two settings, so that one machine may show results considerably higher than another. Likewise, even if two readers of radiographs have similar discriminative power, as shown by similar ROCs, the threshold they use to differentiate positive from negative tests (or adjacent categories of a multicategory test) may vary widely.[37–40]

In summary, variability in the discriminative power and calibration of the same test used in different places is the rule rather than the exception.

When we strive for parsimony in our descriptions, we run the risk of oversimplification. In the end, the researcher who reports a study, as well as the clinician searching the literature for help in interpreting test results, has to bear in mind that test performance characteristics are never just properties of the test itself: they depend on several factors, including prior clinical and test information, and the setting in which the test is done.

2. Implications of variation in discriminative power and calibration of tests: questions to ask yourself before you start designing the study

1 What is the target condition and the reference standard?
2 Is the objective to estimate test performance using a global measure, or a measure that will allow estimation of the probability of disease in individuals?
3 What is the population and clinical problem?
4 Is the test being considered as a replacement or incremental test?
5 To what extent do you want to study the reasons for variability of the results within your population?
6 To what extent do you want to study the transferability of the results to other settings?

In what follows, we assume that the usual criteria for adequate design of an evaluation of a diagnostic test have been fulfilled. The issue is then: How do we design a study which will also help to ensure that its transferability can be determined? Based on the concepts in the first part of this chapter, we suggest that investigators ask themselves the following questions to help ensure that readers have the necessary information to decide on the transferability of the study to their own setting.

1. What is the target condition and reference standard?

The target condition and reference standard need to be chosen to reflect the investigator's requirements. Is the choice appropriate to whether the investigator is doing the study to assist with predicting prognosis, deciding as the need for intervention, or researching pathological processes? For example, in a study of tests to assess stenosis of the carotid artery, it would be sensible to choose the reference standard as angiographic stenosis dichotomised around 70%, if this is the level of angiographic abnormality above which, on currently available evidence, the benefits of treatment outweigh the harm. On the other hand, if the study is being done by researchers whose interest is in basic science, they may wish to compare the test with stenosis assessed on surgically removed specimens at a different threshold, or across a range of thresholds.

Error in the reference standard is a major constraint on our ability to estimate test accuracy and explore reasons for variability of test characteristics.[18,19,41] Therefore, researchers should consider methods of minimising error in the reference standard, for example by using better methods or multiple assessments. Any information about the test performance characteristics of the reference standard will help interpretation, as will several different measures of the target condition, which can be combined. Multiple measures of the reference standard or multiple different tests also allow the use of more sophisticated analyses, such as latent class analysis, to minimise the potential for bias in estimates of test accuracy or factors that affect it.[21,42] Because the effects of misclassification in the reference standard have different effects in populations of different prevalence, as shown in Table 6.2, one may choose to assess a test in a population where any residual effects of error in the reference standard are minimised. For the odds ratio, this is at about 50% prevalence. For sensitivity it is when prevalence is high, and for specificity when prevalence is low. However, when using this strategy, consider whether the spectrum of disease may also vary with prevalence. If so, you will need to judge whether reference standard misclassification is a sufficiently important problem to outweigh the potential for spectrum bias induced by choosing a study in a population with specified prevalence.

2. Is the objective to estimate test performance using a global measure (discrimination) or a measure that will allow estimation of the probability of disease in individuals (discrimination and calibration)?

Global assessment of the discriminatory power of the test requires measures such as the area under the ROC curve, or the diagnostic odds ratio. These may be sufficient for some purposes, for example if a policy decision needs to be made about alternative tests of equivalent cost, or to decide whenever a test has sufficient accuracy to warrant further calibration. For estimating the probability of disease in individuals, likelihood ratios (or sensitivity and specificity) are needed, with additional information on how tests were calibrated. Information about calibration should be provided in papers for readers to be able to use the result of your study. Access to selected example material, such as radiographs of lesions, will help readers understand what thresholds have been used for reading in your study.

3. What is the population and clinical problem?

This question defines how the inception cohort should be selected for study, although the breadth of the group selected will also be determined by the extent to which you wish to address the following questions.

For example, a new test for carotid stenosis could be considered for all patients referred to a surgical unit. However, ultrasound is reasonably accurate at quantifying the extent of stenosis, and so investigators may choose to restrict the study of a more expensive or invasive test to patients in whom the ultrasound result is near the decision threshold for surgery. A useful planning tool is to draw a flow diagram of how patients reach the population/clinical problem of interest. This flow diagram includes what clinical information has been gathered and what tests have been done, and how the results of those tests determine entry into the population and clinical problem of interest. For example, in the flow diagram in Figure 6.4 the clinical problem is suspected appendicitis in children presenting to a hospital emergency service. The decisions based sequentially on clinical evidence and ultrasonography are shown. The flow diagram helps to clarify that computed tomography (CT) is being assessed only in patients in whom those prior tests had not resolved the clinical problem. Also as shown in the figure, in addition to being helpful at the design stage, publishing

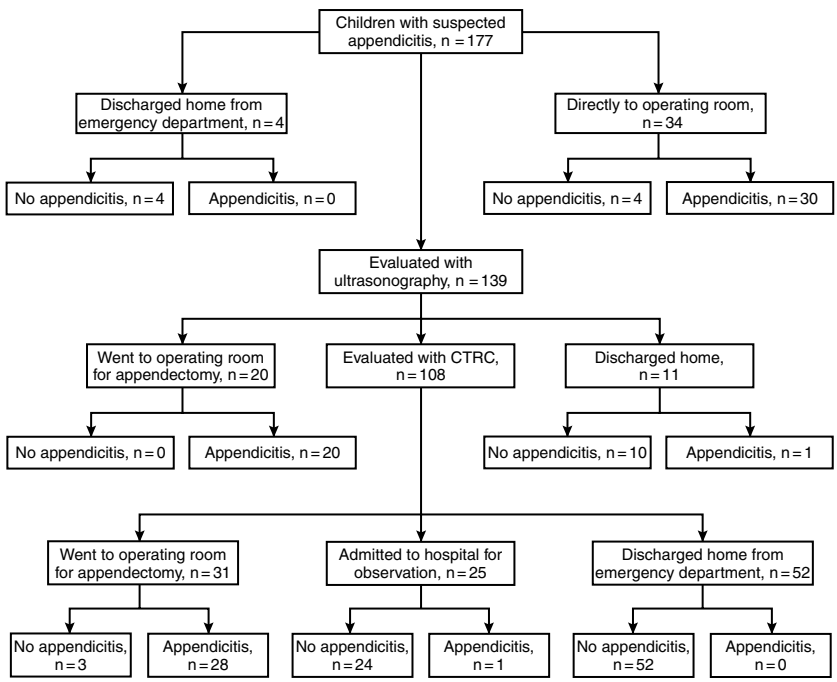


Figure 6.4 A flow diagram to formulate a diagnostic test research question. Study profile flow diagram of patients with suspected appendicitis. (From Garcia Pena BM *et al*. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA* 1999;**282**:1041–6. Reproduced with permission from the American Medical Association.[43])

such flow diagrams, with numbers of patients who follow each step, is very helpful to readers.[43]

4. Is the test being considered as a replacement or incremental test?

As outlined above, the population and the clinical problem define the initial presentation and referral filter. In addition, a key question is whether we are evaluating the test to assess whether it should replace an existing test (because it is better, or just as good and cheaper) or to assess whether it has value when used in addition to a particular existing test. This decision will also be a major determinant of how the data will be analysed.[44–46]

5. To what extent do you want to study the reasons for variability of the results within your population?

How much variability is there between readers/operators?

Data should be presented on the amount of variability between different readers or test types and tools to help calibration, such as standard radiographs,[39,40] or laboratory quality control measures. The extent to which other factors, such as experience or training, affect reading adequacy will also help guide readers of the study. Assessment of variability should include not only test discriminatory power but also calibration, if the objective is to provide study results that are useful for individual clinical decision making.

Do the findings vary in different (prespecified) subgroups within the study population?

Data should be analysed to determine the influence on test performance characteristics of the following variables, which should be available for each individual.

- The spectrum of disease and non-disease, for example by estimating "specificity" within each category of "non-disease". These can be considered separately by users or combined into a weighted specificity for different settings. The same approach can be used for levels (stage, grade) in the "diseased" group.
- The effect of other test results. This follows the approach often used in clinical prediction rules. It should take account of logical sequencing of tests (simplest, least invasive, and cheapest are generally first). It should also take account of possible effect modification by other tests. In some instances people would have been referred because of other tests being positive (or negative), so that the incremental value of the new test cannot be evaluated. In this case, knowing the referral filter and how tests have

been used in it (as in Figure 6.4) will help interpretation. For example, a study by Flamen[31] has shown that the major value of PET for recurrent colorectal adenocarcinoma is in the category of patients in whom prior (cheaper) tests gave inconclusive results. It would therefore be a useful incremental test in that category of patients, but would add little (except cost) if being considered as a replacement test for all patients, many of whom would have the diagnostic question resolved by the cheaper test. This suggests that PET is very helpful in this clinical situation.[31]

- Any other characteristics, such as age or gender.

There are often a vast number of characteristics that could be used to define subgroups in which one may wish to check whether there are differences in test performance. The essential descriptors of a clinical situation need to be decided by the researcher. As for subgroup analysis in randomised trials,[47] these characteristics should be prespecified, rather than decided at analysis stage. The decision is best made on the basis of an understanding of the pathophysiology of the disease, the mechanism by which the test assesses abnormality, an understanding of possible referral filters, and knowledge of which characteristics vary widely between centres. Remember that variability between test characteristics in subgroups may not be due to real subgroup differences if there is reference standard misclassification and the prevalence of disease differs between subgroups, as shown in Table 6.2. Modelling techniques can be used to assess the effect of several potential predictors of test accuracy simultaneously.[48–52]

6. To what extent do you want to study the transferability of the results to other settings?

To address this question, you need to perform the study in several populations or centres, and assess the extent to which test performance differs, as has been done for the General Health Questionnaire[53] and predictors of coma.[54] The extent to which observed variability is beyond that compatible with random sampling variability can be assessed using statistical tests for heterogeneity. Predictors (as discussed above) should also be measured to assess the extent to which within-population variables explain between-population variability. Because of the low power of tests of heterogeneity, this is worth doing even if tests for heterogeneity between centres or studies are not statistically significant. The more the measured variables explain between-population differences, the more they can be relied on when assessing the transferability of that study to the population in the reader's setting. Between-site variability can also be explored across different studies using meta-analytical techniques.[55,56]

Sites for inclusion in the multicentre comparison should be selected as being representative of the sorts of populations in which the results of the diagnostic study are likely to be used. The more the variability in site

features can be characterised – and indeed taken account of in the sampling of sites for inclusion in studies – the more informative the study will be. Data should be analysed to determine the influence on results of the within-site (individually measured) patient characteristics mentioned above. They should also explore the following sources of between-site variability that are not accounted for by the within-site characteristics:

- Site characteristics, for example primary, secondary or tertiary care
- Other features, such as country
- Prevalence of the disease of interest.

Residual heterogeneity between sites should be explored to judge the extent to which there is inexplicable variability that may limit test applicability.

Explanatory note about prevalence

The inclusion of "prevalence" in the above list may seem unusual, as it is not obviously a predictor of test performance. However, there are many reasons why prevalence should be included in the list of potential predictors, in an analogous way to the exploration of trial result dependence on baseline risk.[57] First, many of the reasons for variation between centres may not be easy to characterise, and prevalence may contain some information about how centres differ that is not captured by other crude information, for example whether the test is evaluated in primary, secondary or tertiary care centres. Second, it is a direct test of the common assumption that test performance characteristics such as sensitivity and specificity are independent of prevalence. Third, non-linear prevalence dependence is an indication that there is misclassification of the reference standard.

In summary, there is merit in designing studies with heterogeneous study populations. This will allow exploration of the extent to which diagnostic performance depends on prespecified predictors, and how much residual

Table 6.8 The value of designing studies that enable the exploration of predictors of heterogeneity of diagnostic accuracy.

Heterogeneity in diagnostic accuracy	Heterogeneity in study population	
	Yes	No
Yes	To what extent is heterogeniety in accuracy explained by predictors? If not, transferability is limited	Transferability limited
No	Highly transferable	Design does not allow exploration of transferability

heterogeneity exists. The more heterogeneity there is in study populations, the greater the potential to explore the transferability of diagnostic performance to other settings, as shown in Table 6.8.

Concluding remarks

There is good evidence that measures of test accuracy are not as transferable across settings as is often assumed. This chapter outlines the conceptual underpinnings for this, and suggests some implications for how we should be designing studies that carry more information about the transferability of results. Major examples are examining the extent to which test discrimination and calibration depend on prespecified variables, and the extent to which there is residual variability between study populations which is not explained by these variables. This will require larger studies than have generally been done in the past for diagnostic tests. Improvements in study quality and designs to assess transferability are needed to ensure that the next generation of studies on test accuracy are more able to meet our needs.

Acknowledgement

We thank Petra Macaskill, Clement Loy, André Knottnerus, Margaret Pepe, Jonathan Craig, and Anthony Grabs for comments on earlier drafts. Clement Loy also provided Figure 6.3. We thank Barbara Garcia Pena for permission to use a figure from her paper as our Figure 6.4.

References

1 Begg C. Biases in the assessment of diagnostic tests. *Stat Med* 1987;**6**:411–23.
2 Moons K, van Es G, Deckers JW, Habbema JO, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: A clinical example. *Epidemiology* 1997; **8**:12–17.
3 Detrano R, Janosi A, Lyons KP *et al*. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. *Am J Med* 1988;**84**:699–710.
4 Hlatky M, Pryor D, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med* 1984;**77**:64–71.
5 Rozanski ADG, Berman D, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983;**309**:518–22.
6 Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;**117**:135–40.
7 Molarius ASJ, Sans S, Tuomilehto J, Kuulasmaa K. Varying sensitivity of waist action levels to identify subjects with overweight or obesity in 19 populations of the WHO MONICA Project. *J Clin Epidemiol* 1999;**52**:1213–24.
8 Starmans R, Muris JW, Fijten GH, Schouten HJ, Pop P, Knottnerus JA. The diagnostic value of scoring models for organic and non-organic gastrointestinal disease, including the irritable-bowel syndrome. *Med Decision Making* 1994;**14**:208–16.

9 Glasziou P, Irwig L. An evidence-based approach to individualising treatment. *BMJ* 1995;**311**:1356–9.
10 National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. Handbook series on preparing clinical practice guidelines. Canberra, Commonwealth of Australia, 2000; (Website http://www.health.gov.au/nhmrc/publicat/synopses/cp65syn.htm)
11 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515–24.
12 Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: experience of the Radiology Diagnostic Oncology Group. *Radiology* 1990;**175**:571–5.
13 Sackett D, Straus S, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based Medicine: How to practice and teach EBM*. Edinburgh: Churchill Livingstone, 2000.
14 Richardson W, Wilson M, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club* 1995;**123**:A12.
15 Brenner H. Measures of differential diagnostic value of diagnostic procedures. *J Clin Epidemiol* 1996;**49**:1435–9.
16 Sappenfield RW, Beeler MF, Catrou PG, Boudreau DA. Nine-cell diagnostic decision matrix. A model of the diagnostic process: a framework for evaluating diagnostic protocols. *Am J Clin Pathol* 1981;**75**:769–72.
17 Taube A, Tholander B. Over- and underestimation of the sensitivity of a diagnostic malignancy test due to various selections of the study population. *ACTA Oncol* 1990; **29**:1–5.
18 Buck A, Gart J. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol* 1966;**83**:586–92.
19 Gart J, Buck A. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Am J Epidemiol* 1966;**83**:593–602.
20 Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in observational epidemiology*, 2nd edn. New York: Oxford University Press, 1996.
21 Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;**52**:943–51.
22 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.
23 Coley C, Barry M, Fleming C, Mulley AG. Early detection of prostate cancer. Part I: Prior probability and effectiveness of tests. *Ann Intern Med* 1997;**126**:394–406.
24 van Beek EJR, Schenk BE, Michel BC. The role of plasma D-dimer concentration in the exclusion of pulmonary embolism. *Br J Haematol* 1996;**92**:725–32.
25 Knottnerus J, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;**45**:1143–54.
26 van der Schouw YT, van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol* 1995;**48**:417–22.
27 Katz IA, Irwig L, Vinen JD, *et al*. Biochemical markers of acute myocardial infarction: strategies for improving their clinical usefulness. *Ann Clin Biochem* 1998;**35**:393–9.
28 Whitsel E, Boyko E, Siscovick DS. Reassessing the role of QT in the diagnosis of autonomic failure among patients with diabetes: a meta-analysis. *Arch Intern Med* 2000;**23**:241–7.
29 Conde-Agudelo A, Kafury-Goeta AC. Triple-marker test as screening for Down's syndrome: a meta-analysis. *Obstet Gynecol Surv* 1998;**53**:369–76.
30 Mol B, Hajenius P, Engelsbel S. Serum human chorionic gonadotrophin measurement in the diagnosis of ectopic pregnancy when transvaginal sonography is inconclusive. *Fertil Steril* 1995;**70**:972–81.
31 Flamen PSS, van Cutsem E, Dupont P, *et al*. Additional value of whole-body positron emission tomography with fluorine-18-2-fluoro-2-deoxy-D-glucose in recurrent colorectal cancer. *J Clin Oncol* 1999;**17**:894–901.
32 Decode Study Group. Glucose tolerance and mortality: comparison of WHO and American Diabetes Association diagnostic criteria. *Lancet* 1999;**354**:617–21.

33 Scouller K, Conigrave KM, Macaskill P, Irwig L, Whitfield JB. Should we use CDT instead of GGT for detecting problem drinkers? A systematic review and meta-analysis. *Clin Chem* 2000;**46**:1894–902.
34 West OC, Anbari MM, Pilgram TK, Wilson AJ. Acute cervical spine trauma: diagnostic performance of single-view versus three-view radiographic screening. *Radiology* 1997;**204**:819–23.
35 Scheiber CMM, Dumitresco B, Demangeat JL, *et al*. The pitfalls of planar three-phase bone scintigraphy in nontraumatic hip avascular osteonecrosis. *Clin Nucl Med* 1999;**24**:488–94.
36 Krupinski EA, Weinstein RS, Rozek LS. Experience-related differences in diagnosis from medical images displayed on monitors. *Telemed J* 1996;**2**:101–8.
37 Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. *JAMA* 1996;**276**:1752–5.
38 Irwig L, Groeneveld HT, Pretorius JP, Itnizdo E. Relative observer accuracy for dichotomized variables. *J Chronic Dis* 1985;**28**:899–906.
39 D'Orsi C, Swets J. Variability in the interpretation of mammograms. *N Engl J Med* 1995;**332**:1172.
40 Beam C, Layde P, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med* 1996;**156**:209–13.
41 Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol* 1990;**93**:252–8.
42 Walter S, Irwig L. Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *J Clin Epidemiol* 1988;**41**:923–37.
43 Garcia Pena BM, Mandl KD, Kraus SJ, *et al*. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA* 1999;**282**:1041–6.
44 Marshall R. The predictive value of simple rules for combining two diagnostic tests. *Biometrics* 1989;**45**:1213–22.
45 Biggerstaff B. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics Med* 2000;**19**:649–63.
46 Chock C, Irwig L, Berry G, Glaszion P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;**50**:1211–17.
47 Oxman A, Guyatt G. A consumer guide to subgroup analyses. *Ann Intern Med* 1992;**116**:78–84.
48 Tosteson ANA, Weinstein MC, *et al*. ROC curve regression analysis: the use of ordinal regression models for diagnostic test assessment. *Environ Health Perspect* 1994;**102**(Suppl 8):73–8.
49 Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics Med* 1996;**15**:1807–26.
50 Pepe M. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 2000;**56**:352–9.
51 Leisenring W, Pepe M. Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics* 1998;**54**:444–52.
52 Leisenring W, Pepe M, Longton G. A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics Med* 1997;**16**:1263–81.
53 Furukawa T, Goldberg DP. Cultural invariance of likelihood ratios for the General Health Questionnaire. *Lancet* 1999;**353**:561–2.
54 Zandbergen EGJ, de Haan RJ, *et al*. Prognostic predictors of coma transferable from one setting to another in SR. *Lancet* 1998;**352**:1808–12.
55 Irwig L, Tosteson ANA, Gatsonis C, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.
56 Rutter C, Gatsonis C. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;**2**(Suppl 1):S48–56.
57 Schmid C, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics Med* 1998;**17**:1923–42.

7 Analysis of data on the accuracy of diagnostic tests

J DIK F HABBEMA, RENÉ EIJKEMANS,
PIETA KRIJNEN, J ANDRÉ KNOTTNERUS

Summary box

- Neither sensitivity nor specificity is a measure of test performance on its own. It is the combination that matters.
- The statistical approach for analysing variability in probability estimates of test accuracy is the calculation of confidence intervals.
- The magnitude of the change from pretest to post-test probability (predictive value) reflects the informativeness of the diagnostic test result.
- The informative value of a test result is determined by the likelihood ratio: the ratio of the frequencies of occurrence of this result in patients with and patients without the disease.
- The odds ratio summarises the diagnostic value of a dichotomous test, but does not tell us the specific values of sensitivity and specificity and the likelihood ratios.
- A measure of performance for a continuous test is the area under the ROC curve. This varies between 0.5 for a totally uninformative test and 1.0 for a test perfectly separating diseased and non-diseased.
- Bayes' theorem implies that "post-test odds equals pretest odds times likelihood ratio".
- One can derive the optimal cut-off from the relative importance of false positives and false negatives.
- A sensitivity analysis is important for getting a feeling for the stability of our conclusions.

- From multiple logistic regression analysis one can not only learn about the predictive value of a combination of tests, but also what a certain test adds to other tests that have already been performed.
- When starting a data analysis one must be confident that the research data have been collected with avoidance of important bias and with acceptable generalisability to the target population.

Introduction

After the painstaking job of collecting, computerising and cleaning diagnostic data, we enter the exciting phase of analysing and interpreting these data and assessing the clinical implications of the results. It would be a pity if all the effort put into the research were not to be crowned with a sound analysis and interpretation. It is the purpose of this chapter to help readers to do so.

We will study the classic test performance measures introduced in Chapter 1: sensitivity, specificity, positive and negative predictive value, likelihood ratio, and error rate, first for dichotomous tests and later, for continuous tests, including the possibility of dichotomisation, with its quest for cut-off values. ROC curves are part of this.

Next, Bayes' theorem for the relationship between pretest and post-test probability of disease is discussed, followed by decision analytical considerations. For generalisation of the one-test situation to diagnostic conclusions based on many diagnostic test results, there will be a discussion on logistic regression and its link with Bayes' theorem.

The strengths and weaknesses of study designs, possible biases, and other methodological issues have been discussed in previous chapters and will not be repeated here, although the discussion will provide some links between biases and analysis results.

We will refer to software for performing the analysis. Also, we will include appendices with tables and graphs, which can support you in the analysis.

Clinical example

Renal artery stenosis in hypertension

We use data from a study on the diagnosis of renal artery stenosis. In about 1% of all hypertensive patients the hypertension is caused by a constriction (stenosis) of the renal artery. It is worth identifying these patients because their hypertension could be cured by surgery, and consequently their risk of myocardial infarction and stroke could be

Table 7.1 The first three and last three patients of 8×437 data array from a study on diagnostics in possible renal artery stenosis (RAS).

Patient code	Age	Gender	Atherosclerotic vascular disease	Abdominal bruit	Creatinine (micromol)	Abnormal renogram	RAS on angiography
1	62	F	No	Yes	87	No	Yes
2	52	M	No	No	146	Yes	Yes
3	49	F	No	No	77	No	No
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
435	36	M	No	No	84	No	No
436	51	M	Yes	No	74	No	No
437	55	M	No	No	83	No	No

reduced. Moreover, renal failure could be prevented by relieving the stenosis. The definitive diagnosis of renal artery stenosis is made by renal angiography. This diagnostic reference test should be used selectively, because it is a costly procedure that can involve serious complications. Thus, clinicians need a safe, reliable, and inexpensive screening test to help them select patients for angiography.

The diagnostic tests that we will use in this chapter are clinical characteristics suggestive of renal artery stenosis, and renography; angiography serves as the reference standard test for stenosis. The clinical characteristics used as examples are symptoms and signs of atherosclerotic vascular disease, the presence of an abdominal bruit and the serum creatinine concentration. Renography is a non-invasive test for detecting asymmetry in renal function between the kidneys, which also is suggestive of renal artery stenosis.

The data, listed as indicated in Table 7.1, are from a Dutch multicentre study aiming to optimise the diagnosis and treatment of renal artery stenosis (RAS). The study included 437 hypertensive patients aged 18–75 years, who had been referred for unsatisfactory blood pressure control or for analysis of possible secondary hypertension.

Diagnostic questions and concepts

One can ask a number of questions concerning this diagnostic problem. Some are mentioned below, with the diagnostic concept concerned in parentheses.

- How good is my diagnostic test in detecting patients with RAS (sensitivity)?
- How good is my diagnostic test in detecting patients without RAS (specificity)?
- How well does a positive/abnormal test result predict the presence of RAS (positive predictive value)?

- How well does a negative/normal test result predict the absence of RAS (negative predictive value)?
- What is a reasonable estimate for the pretest probability of RAS (prevalence of RAS)?
- How many false conclusions will I make when applying the diagnostic test (error rate)?
- How informative is my positive/negative test result (likelihood ratio)?
- How do I summarise the association between a dichotomous test and the standard diagnosis (diagnostic odds ratio)?
- What is an optimal cut-off level when I want to dichotomise a continuous test (ROC curve)?
- To what extent does the test result change my pretest belief/probability of RAS (Bayes' theorem)?
- How are the above concepts applied to a number of diagnostic tests simultaneously (logistic regression)?

Sensitivity and specificity for a dichotomous test

We will illustrate the dichotomous test situation by assessing how well renography is able to predict arterial stenosis. Therefore we construct from our database the 2×2 table with, as entries for renographic assessment, "abnormal/normal", and for angiography, "stenosis/no stenosis" (Table 7.2).

The generic table with the corresponding symbolism is given in Table 7.3, with N = total number of patients, N_{T+} = number of patients with positive test results, N_{T-} = number of patients with negative test results,

Table 7.2 2×2 table for analysing the diagnostic value of renographic assessment in predicting renal artery stenosis.

	Angiography		
Renography	Stenosis	No Stenosis	Total
Abnormal	71	33	104
Normal	29	304	333
	100	337	437

Table 7.3 Generic 2×2 table representing possible classifications for the relationship between a diagnostic test and a diagnosis (reference test).

Test	Diagnosis		
	+	−	
+	TP	FP	N_{T+}
−	FN	TN	N_{T-}
	N_{D+}	N_{D-}	N

N_{D-} = number of patients without the disease, N_{D+} = the number of patients with the disease, TP = number of true positives, TN = number of true negatives, FP = number of false positives, and FN = number of false negatives.

Together, *sensitivity* (the probability of a positive test result in diseased subjects, $P(T+|D+)$ and *specificity* (the probability of a negative test result in non-diseased subjects, $P(T-|D-)$, characterise the dichotomous test for the clinical situation at hand. Neither is a measure of test performance on its own: it is the combination that matters.[2]

For the example in Table 7.2 we can calculate:

Sensitivity = $TP/N_{D+} = 71/100 = 71\%$
Specificity = $TN/N_{D-} = 304/337 = 90\%$

In the next section we will see the degree of variability with which these estimates are associated.

Sampling variability and confidence intervals for probabilities

Confidence intervals

A main challenge in the analysis of diagnostic data is to assess how confident we can be about the test characteristics as observed in our patients. This may sound strange because an observed proportion, for example the sensitivity of renography in Table 7.2 of 71%, is a fact. However, it is unlikely that we will again find exactly 71% for sensitivity in a new series of 437 similar patients, and an indication of the limits of what can reasonably be expected is therefore important (even when the same patients would have been re-examined in the same or another setting, other data will be obtained because of inter- and intraobserver variability).

The statistical method for analysing the variability in estimates of sensitivity, and of all other probability estimates that we will discuss, is confidence intervals. The probability level of the confidence interval can be chosen. A higher level of confidence corresponds to a larger interval in terms of number of percentiles covered. Throughout the chapter, we will – conventionally – work with 95% confidence intervals. The interpretation of a *95% confidence interval* for an observed proportion, that is, a probability estimate, is as follows: when the data sampling is repeated many times, the 95% confidence interval calculated from each sample will, on average, contain the "true" value of the proportion in 95% of the samples. Variability in sensitivity estimates is illustrated in Table 7.4. In part(a) of this table, the 100 stenosis patients of our study are subdivided into four groups of 25 consecutive patients. It is seen that the four subgroup sensitivities range enormously, from 48% to 88% (tables and formulae for the confidence interval will be discussed later). Part(b) illustrates what sensitivities we would

have obtained if we had finished the study earlier, that is, after observing the first 5, 10, 25, 50 and 100 stenosis patients of the present study.

As you see from Table 7.4(a), the 95% confidence intervals of the highest and lowest estimates of sensitivity of 0.88 and 0.48 just touch each other. Table 7.4(b) shows that the width and the confidence interval become smaller with increasing sample size, as you would expect. For confidence intervals, and more generally for the accuracy of statistical estimates, the square root rule applies: when one makes the sample size A times as large, the confidence interval will be a factor $\sqrt{A}$ smaller. For example, for a two-times smaller confidence interval one needs four times as many patients. You can check the (approximate) validity of the square root rule in Table 7.4(b).

The confidence interval for the 71% sensitivity estimate for the total study runs from 61% to 80% (bottom line in Table 7.4). Table 7.5 gives confidence intervals for a number of confidence levels, with wider intervals for higher levels.

For the specificity of renography in diagnosing RAS we get the following confidence interval around the 90% estimate for the total number of 337 non-diseased subjects: from 87% to 93%. As you can see, the confidence

Table 7.4 Analysis of diagnostic data of patients with possible renal artery stenosis (RAS): confidence intervals for sensitivity of renography in diagnosing RAS; (a) variability in sensitivity between equal numbers of RAS patients, and (b) smaller confidence intervals with larger sample size, as cumulated during the study.

	TP	N_{D+}	Sensitivity	95% Confidence interval
(a)	18	25	0.72	0.51–0.88
	22	25	0.88	0.69–0.97
	19	25	0.76	0.55–0.91
	12	25	0.48	0.28–0.69
(b)	3	5	0.60	0.15–0.95
	7	10	0.70	0.35–0.93
	18	25	0.72	0.51–0.88
	40	50	0.80	0.66–0.90
	71	100	0.71	0.61–0.80

Table 7.5 Confidence interval for the 71% (71 out of 100) sensitivity estimate of renography in diagnosing RAS, for different confidence levels.

Confidence level (%)	Confidence interval (%)
50	67–74
67	66–76
80	64–77
90	63–78
95	61–80
99	58–82
99.9	54–84

interval is roughly half the size of the confidence interval for the sensitivity, which reflects about four times as high the number of observations on which the estimate is based (square root rule!).

Some theory and a guide to the tables in the appendix

The theory of calculating confidence intervals for proportions is based on the binomial distribution and requires complicated calculations. In general, the confidence interval is asymmetrical around the point estimate of the sensitivity, because of the "floor" and "ceiling" effects implied by the limits of 0 and 1 to any probability.

Fortunately, when the numbers are not small the 95% confidence interval becomes approximately symmetrical and the upper, and lower limits can be calculated by adding or substracting $1.96 \times$ *standard error*, with the standard error calculated by:

$$\sqrt{\hat{p}(1 - \hat{p})/N}$$

where $\hat{p}$ stands for the proportion or probability estimate, and N for the number of observations on which the proportion is based (in practice, multiplication by 2 instead of the more tedious 1.96 works well).

For other confidence levels, the multiplication factor 1.96 should be replaced by other values (see Appendix A.3).

Appendix Tables A.1 and A.2 give confidence levels for situations with small sample sizes where you need the tedious binomial calculations. Table A.3 gives the confidence interval for a number of situations in which the above formula for the standard error works well. In cases not covered by the tables, the standard error can be calculated using the formula. The reader can now verify the correctness of the confidence intervals presented in this section.

Positive and negative predictive value: pre- and post-test probability of disease

The *positive predictive value* (PPV) is the probability that the patient has the disease when the test result is positive. This "post-test probability" is easily derived from Table 7.2. For the probability of RAS in case of abnormal renography, it is:

$$\text{PPV} = \text{P}(\text{D}+ \mid \text{T}+) = \text{TP}/\text{N}_{\text{T}+} = 71/104 = 68\%$$

The confidence interval (CI) can be estimated using the formula on page 137 or Table A.3.: the 95% confidence interval for PPV runs from 59% to 77%.

The *negative predictive value* (NPV), that is, the probability that the patient has the disease if the test result is negative, translates in our case to

the probability of no stenosis in case of normal renography. We get:

$$NPV = P(D-|T-) = TN/N_{T-} = 304/333 = 91\%$$

with 95% CI from 88% to 94% (see Table A.3).

The probabilities of no stenosis for an abnormal renogram and of stenosis for a normal renogram, and their CIs, are obtained as 100% minus PPV and 100% minus NPV, respectively because the probabilities of stenosis and no stenosis have to add up to 100%:

$$P(D-|T+) = 1-P(D+|T+) = 32\% \text{ (95\% CI: 23\% to 41\%)}$$
$$P(D+|T-) = 1-P(D-|T-) = 9\% \text{ (95\% CI: 6\% to 12\%)}$$

The PPV and NPV are *post-test* probabilities, that is, they are the updated probabilities given the information provided by the positive and negative test results, respectively. Before the test, we have the *pretest* probabilities of presence and absence of disease, which for our RAS example are:

$$P(D+) = 100/437 = 23\% \text{ (95\% CI: 19\% to 27\%)}$$
$$P(D-) = 337/437 = 77\% \text{ (95\% CI: 73\% to 81\%)}$$

The magnitude of the change from pre- to post-test probability reflects the informativeness of the diagnostic test result. In our case, the pretest probability of stenosis of 23% changes to 68% in case of an abnormal renogram, and to 9% in case of a normal renogram.

Error rate

How well does our diagnostic test discriminate between patients with and without stenosis, or, more generally, how well does the test discriminate between the two disease categories? So far we have only looked at partial measures of performance, such as sensitivity, specificity, and predictive values. None of these concepts on its own gives an assessment of the performance of the test.

The most straightforward measure expresses how many errors we make when we diagnose patients with an abnormal test result as diseased, and those with a normal test result as non-diseased. This concept is known as the *error rate*. For our example, the error rate is easily calculated from Table 7.1. There are 29 false negative results, as the test was negative when stenosis was present, and 33 false positive results, with the test being abnormal when there was no stenosis. Thus, in total there are 62 errors, from a total of 437 patients.

This gives the following calculations for the error rate and its confidence interval, the latter being derived from Table A.3 (the closest entry is 60 out

of 500, with a half-CI size of 0.028; interpolation to 70 and 300 shows that ± 3% is indeed the correct CI):

Error rate(ER) = 62/437 = 14% (95% CI: 11% to 17%)

The error rate is a weighted average of errors among persons with the disease (the false negatives) and among those without the disease (the false positives), as is seen from the following equation:

$$ER = P(T-|D+) \times P(D+) + P(T+|D-) \times P(D-)$$

For our stenosis example we can easily verify this expression for the error rate:

$$ER = (29/100) \times (100/437) + (33/337) \times (337/437) = 62/437 = 14\%$$

The weights in this formula are 23%, being the pretest probability of disease, and 77% for the probability of no disease.

This equation enables us to investigate what the error rate would be if the pretest probability of disease were different. For example, if the pretest probability of disease were 50% instead of 23%, the error rate would be calculated as:

$$ER = 29/100 \times 0.5 + 33/337 \times 0.5 = 19.4\%$$

Using this formula we can speculate about the performance of the test in situations that differ from the original context (the assumption is that false positive and false negative rates do not change. This is unfortunately not always valid; see Chapters 1, 2, and 6).

Information in a diagnostic test result: the likelihood ratio

The informative value, or weight of evidence, of a test result is determined by the frequency of occurrence of this result in patients with the disease compared to those without the disease. If, for example, a certain test result occurs twice as often in patients with the disease, this result gives an evidence factor of 2 in favour of the disease. If, on the other hand, a test result occurs twice as often in patients without the disease, it gives an evidence factor of 2 in favour of non-disease, that is, a factor 2 against the disease (or a factor 1/2 in favour of disease).

This important probability ratio is called the likelihood ratio (LR). Each test result X has its own likelihood ratio $LR(X) = P(X|D+)/P(X|D-)$.

For dichotomous tests, we have only two test results, T+ and T−, and therefore also only two likelihood ratios:

the LR of a positive test result: $LR(T+) = P(T+|D+)/P(T+|D-) = Se/(1-Sp)$

the LR of a negative test result: $LR(T-) = P(T-|D+)/P(T-|D-)$
$= (1-Se)/Sp$

For our example of renal artery stenosis, we obtain the following values for the likelihood ratio of an abnormal and a normal renogram, respectively:

$LR(T+) = 0.71/0.10 = 7.1$ with 95% CI: 5.1 to 10.3
$LR(T-) = 0.29/0.90 = 0.32$ with 95% CI: 0.24 to 0.44.

Thus, an abnormal renogram provides a factor of 7 in favour of stenosis, whereas a normal renogram yields a factor of 3 (that is, 1/0.32) in favour of no stenosis.

The following approximate formula has been used to calculate the 95% confidence interval for the likelihood ratio:

$$\exp\left(\ln\frac{p_1}{p_2} \pm 1.96\sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}\right)$$

in which $p_1 = P(X|D+)$ is based on sample size n_1 and $p_2 = P(X|D-)$ on sample size n_2.[3]

Diagnostic odds ratio

For a dichotomous test it is possible to summarise the association between the test and the diagnosis (reference standard) presented in the 2×2 table in one measure: the diagnostic odds ratio (OR), which is equivalent to the cross-product of the table. Looking at the example of renal artery stenosis (Table 7.2):

$OR = (71/33)/(29/304) = (71 \times 304)/(33 \times 29) = 22.6$,
with 95% CI: 12.4 to 41.3

The OR is equivalent to the ratio of LR(T+) and LR(T−), as can be easily checked in the table. The 95% confidence interval of the OR is provided by the software recommended in the references with this chapter.

The advantage of the OR is that it summarises in one figure the diagnostic association in the whole table. However, this summary measure does not tell us the specific values of the likelihood ratios of the two test results, nor those of sensitivity and specificity. These measures have to be calculated as described earlier.

Continuous tests, and their dichotomisation and trichotomisation

Another test for investigating the presence or absence of renal artery stenosis is the serum creatinine concentration. This test has a continuous

range of possible test results. For analysis, results can best be grouped in classes of sufficient size (Table 7.6). Each class has its own evidence for and against stenosis, as expressed in the likelihood ratio.

The theory thus far has concerned only dichotomous tests, but the specific concepts for the dichotomous test situation can be translated into more general concepts for tests with more categories. The probabilities of observing a test result for stenosis and non-stenosis patients are given in Table 7.6. The likelihood ratio is a concept linked to a specific test result, and so also applies to multicategory tests. For example, the likelihood ratio for a test result in the category 61–70 micromol can be calculated as the ratio of the likelihood of this test result in diseased and the likelihood of this result in non-diseased, that is: (4/100)/(36/337) = 0.37. As expected, the likelihood ratio increases with higher serum creatinine levels. The irregularity in this increasing trend in the 81–90 class reflects sampling variation, and not an underlying biological phenomenon.

We will now analyse the relationship between the multicategory test of serum creatinine described in Table 7.6 and its possible simplification to a dichotomous test. *Dichotomisation* can take place at any category boundary. This is done in Table 7.7, which gives in each row the corresponding dichotomous test data. For example, based on a cut-off level of 80 the number of patients with and without stenosis over the value of 80 is 82 and 215, respectively, resulting in a sensitivity of 82% and a specificity of 36%. Likelihood ratios can again be calculated, now for the two results of the dichotomised test. As can be seen, much information is lost by the dichotomisation. All results above and below the threshold are aggregated, and the likelihood ratio after dichotomisation becomes an average of the likelihood ratios of the individual classes above and below this threshold. Also, the question of the choice of the cut-off value is a difficult one, especially

Table 7.6 Probability of test results and diagnostic information of serum creatinine concentration in relation to renal artery stenosis.

Serum creatinine (micromol/l)	Stenosis	No stenosis	All	Likelihood ratio (95% CI)
≤60	1 (1%)	19 (6%)	20 (5%)	0.18 (0.02–1.31)
61–70	4 (4%)	36 (11%)	40 (9%)	0.37 (0.14–1.03)
71–80	13 (13%)	67 (20%)	80 (18%)	0.65 (0.38–1.13)
81–90	12 (12%)	71 (21%)	83 (19%)	0.57 (0.32–1.01)
91–100	17 (17%)	71 (21%)	88 (20%)	0.81 (0.50–1.30)
101–110	15 (15%)	41 (12%)	56 (13%)	1.23 (0.71–2.13)
111–120	7 (7%)	10 (3%)	17 (4%)	2.33 (0.92–6.04)
121–130	9 (9%)	9 (3%)	18 (4%)	3.33 (1.37–8.26)
131–150	11 (11%)	8 (2%)	19 (4%)	4.58 (1.92–11.20)
>150	11 (11%)	5 (1%)	16 (4%)	7.33 (2.64–20.84)
All	100 (100%)	337 (100%)	437 (100%)	

Table 7.7 Probability of test results and diagnostic information of dichotomised serum creatinine concentration values for nine possible cut-offs between high and low values.

Serum creatinine (Micromol)	Stenosis Se	No stenosis 1 − Sp	All	LR+	LR−
>60	99 (99%)	318 (94%)	417 (95%)	1.05	0.18
>70	95 (95%)	282 (84%)	377 (81%)	1.14	0.31
>80	82 (82%)	215 (64%)	297 (68%)	1.29	0.50
>90	70 (70%)	144 (43%)	114 (49%)	1.64	0.52
>100	53 (53%)	73 (22%)	126 (29%)	2.44	0.60
>110	38 (38%)	32 (9%)	70 (16%)	4.00	0.69
>120	31 (31%)	22 (7%)	53 (12%)	4.77	0.74
>130	22 (22%)	13 (4%)	35 (8%)	5.64	0.81
>150	11 (11%)	5 (1%)	16 (4%)	7.33	0.90
Total	100 (100%)	337 (100%)	437 (100%)		

when patients require a different amount of evidence in deciding for or against a certain further action. In our case, it could well be that some patients with a history that more clearly corroborates renal artery stenosis only need limited further evidence in order to decide for surgical intervention, whereas others need high likelihood ratios for the same decision.

The sensitivity–specificity pairs obtained for different cut-off values can be connected in a graph, yielding the so-called *ROC curve* (Figure 7.1). The more the ROC curve moves toward the left upper corner, which represents a perfect dichotomous test with 100% sensitivity and 100% specificity, the better the test is. The steepness of the slope between two adjoining cut-off points represents the likelihood ratio of an observation falling in between these two points. This is shown in Figure 7.2. The likelihood ratios in Figure 7.2 are the same as those in Table 7.6.

A measure of performance for the test is the *area under the ROC curve*.[4] This varies between 0.5 for a totally uninformative test with a likelihood ratio of 1 for all its cut-off values (the diagonal of Figure 7.1), and 1 for a test that perfectly separates diseased and non-diseased (Se=Sp=1.0). The serum creatinine has an area under the curve of 0.70 for differentiating between stenosis and non-stenosis patients. The interpretation of the value of 0.70 is as follows. Consider the hypothetical situation that two patients, drawn randomly from the stenosis patients and the non-stenosis patients respectively, are subjected to the serum creatinine test. If the test results are used to guess which of the two is the stenosis patient, the test will be right 70% of the time. The confidence interval can be calculated using a computer program (see software references).

If a continuous test such as serum creatinine has to be summarised in a few classes for further condensation of the results or for further decision

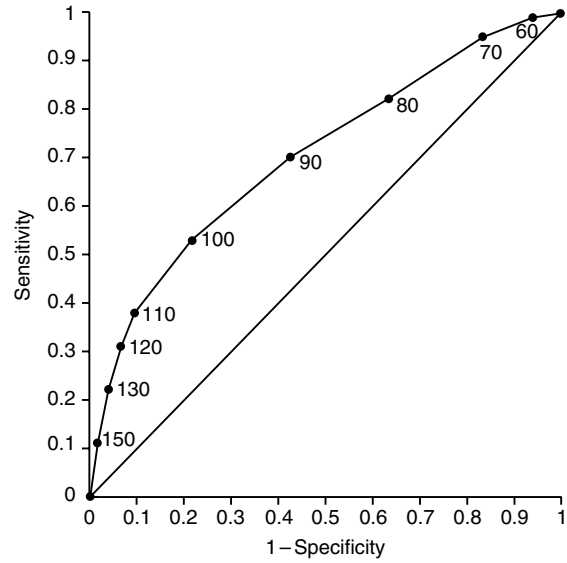


Figure 7.1 Receiver operating characteristic (ROC) curve for serum creatinine concentration in diagnosing renal artery stenosis. For each cut-off value of the serum creatinine, the probability of finding a higher value in stenosis (Se) and in non-stenosis patients (1−Sp) is plotted. The area under the ROC curve is 0.70 (95% CI 0.64–0.76).

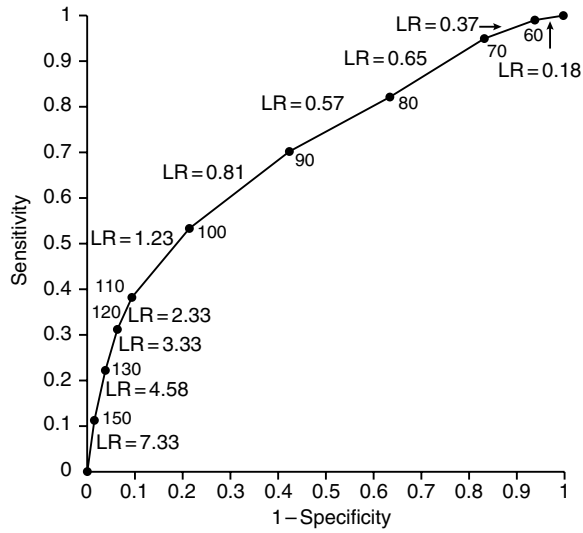


Figure 7.2 Receiver operating characteristic (ROC) curve for serum creatinine concentration in diagnosing renal artery stenosis, with the likelihood ratio for stenosis for each class of serum creatinine values.

Table 7.8 Probability of test results and diagnostic information of serum creatinine concentration for a trichotomisation of the test results.

Serum creatinine (Micromol/l)	Stenosis	No stenosis	All	Likelihood ratio
≤70	5 (5%)	55 (16%)	60 (14%)	0.31
71–110	57 (57%)	250 (74%)	307 (70%)	0.77
>110	38 (38%)	32 (10%)	70 (16%)	4.00
All	100 (100%)	337 (100%)	437 (100%)	

making, it is often more useful to consider a trichotomisation than a dichotomisation. In Table 7.8 we have divided the serum creatinine value into three classes, one for values giving a reasonable evidence for stenosis (likelihood ratio greater than 2.0), one for results giving reasonable evidence against stenosis (likelihood ratio smaller than 0.5), and an intermediate class for rather uninformative test results. It is seen that serum creatinine gives informative test results in about 30% of patients, whereas the test results are rather uninformative in the remaining 70%.

From pretest probability to post-test probability: Bayes' theorem

The formula for calculating how the pretest probability changes under the influence of diagnostic evidence into a post-test probability is known as Bayes' theorem. In words, this is as follows:

> *"If disease was A times more probable than no disease before carrying out a certain test, and if the observed test result is B times as probable in diseased as in non-diseased subjects, then the disease is (A × B) as probable compared to no disease after the test."*

A, B and A × B are respectively the pretest odds, the likelihood ratio, and the post-test odds, and a technical formulation of Bayes' theorem is therefore: "post-test odds equals pretest odds times likelihood ratio"; and in formula: $O(X) = O \times LR(X)$. An example: take the dichotomous renography test (Table 7.2). The pretest odds (A) are 100:337, or 0.30. Assuming a positive test result, the likelihood ratio B equals 7.1. Bayes' theorem tells us now that the post-test odds of disease are $0.30 \times 7.1 = 2.13$. This corresponds to a probability of $2.13/(2.13 + 1) = 0.68$, because of the relationship between probability P and odds O: $O = P/(1-P)$, and therefore $P = O/(1 + O)$.

Another example: take the category 61–70 for serum creatinine (Table 7.6) with a likelihood ratio of 0.37. In the post-test situation, stenosis is $0.30 \times 0.37 = 0.11$ times as probable as no disease. This yields a probability of stenosis of $0.11/1.11 = 0.10$.

The formula of Bayes' theorem for directly calculating the post-test probability is as follows:

$$P(D+|X) = \frac{P(D+) \times P(X|D+)}{P(D+) \times P(X|D+) + P(D-) \times P(X|D-)}$$

For dichotomous test we can express this formula in terms of sensitivity (Se) and specificity (Sp), and positive and negative predictive values (PPV and NPV), as can also be easily derived from the (2 × 2) Tables 7.2 and 7.3:

$$PPV = \frac{P(D+) \times Se}{P(D+) \times Se + P(D-) \times (1 - Sp)} \quad \text{and}$$

$$NPV = \frac{P(D-) \times Sp}{P(D-) \times Sp + P(D+) \times (1 - Se)}$$

Figure 7.3 gives a graphical presentation of Bayes' theorem and enables you to directly calculate the post-test probability from pretest probability and likelihood ratio. The two examples described earlier can be graphically

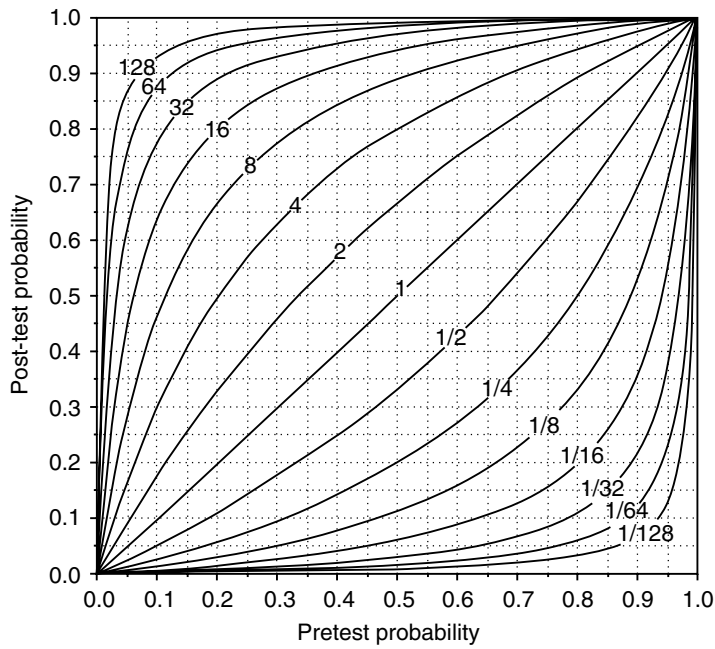


Figure 7.3 Graphical presentation in which the result of Bayes' theorem can be read for a number of likelihood ratios. Each curve represents one likelihood ratio. For example, a pretest probability of 0.7 and a likelihood ratio of 4 give rise to a post-test probability of about 0.90.

verified with a pretest probability of stenosis of 23%: a likelihood ratio of 7.1 gives a post-test probability of 68%, and a likelihood ratio of 0.37 gives a post-test probability of 10% (these post-test probabilities can only be read approximately in the figure). For an alternative, nomogram type of representation of Bayes' theorem, see Chapter 2, Figure 2.1.

Decision analytical approach of the optimal cut-off value

The error rate is a good measure of test performance, as it gives the number of false positives and false negatives in relation to the total number of diagnostic judgements made. It should be realised that the error rate implicitly assumes that false positives and false negatives have an equal weight. This may not be reasonable: for example a missed stenosis may be judged as much more serious than a missed non-stenosis. Here we enter the realm of decision science, where the loss of wrong decisions is explicitly taken into account. As an appetiser to decision analysis, look at Table 7.9. This is based on Table 7.7, but now with an indication of how many false positives are additionally avoided and how many additional false negatives are induced by increasing the threshold between positive and negative ("cut-off") with one class of serum creatinine values at a time.

With the (uninteresting) cut-off value of 0, we would have 337 false positives (FP) and 0 false negatives (FN). Increasing the threshold from 0 to 60 would decrease the FP by 19 and increase the FN by 1. A shift from 60 to 70 would decrease the FP by 36 and increase the FN by four, and so on, until the last step in cut-off from 150 to "very high" serum creatinine

Table 7.9 Decrease in false positives and increase in false negatives when increasing the cut-off between high and low serum creatinine values by one class at a time. Eleven possible cut-offs are considered.

Serum creatinine (micromol/l)	No stenosis	Stenosis	FP decrease: FN increase per step	Approximate trade-off
>0	337	100	337:0	
>60	318	99	19:1	20:1
>70	282	95	36:4	10:1
>80	215	82	67:13	5:1
>90	144	70	71:12	5:1
>100	73	53	71:17	5:1
>110	32	38	41:15	3:1
>120	22	31	10:7	1:1
>130	13	22	9:9	1:1
>150	5	11	8:11	1:1
"Very high"	0	0	5:11	1:2
Total	337	100		

values, by which the last five FP are prevented but also the last 11 stenosis patients are turned into FN.

One can derive the optimal cut-off from the relative importance of false positives and false negatives. For example, if one false positive is judged to be four times more serious than a false negative, a good cut-off would be 100, because all shifts in cut-off between 0 and 100 involve a trade-off of at least five FN to one FP, which is better than the 4 : 1 judgement on the relative seriousness of the two types of error. A further shift from 100 to 110 is not indicated because the associated trade-off is three FN or less to one FP or more, which is worse than the 4 : 1 judgement. Note that for different pretest values of stenosis the FN : FP trade-offs will change, and therefore also the optimal threshold. For example, if the pretest probability were two times higher, the threshold would shift to 60 (calculations not shown).

For a further study of decision analytical considerations, the reader is referred to Sox et al.[5]

Sensitivity analysis

In a sensitivity analysis we look at what would have happened to our conclusions in case of other, but plausible, assumptions. This is important for getting a feeling for the stability of the conclusions.

We saw an example of a sensitivity analysis in our discussion of the error rate, when we looked what the error rate would have been if the pretest probability of stenosis had been different from the 30% in the study.

Sensitivity analysis could also be conducted using Figure 7.3, the graphical representation of Bayes' theorem. Using the confidence intervals for the pretest probability and for the likelihood ratio, we can assess the associated uncertainty in the post-test probability. For example, when we have a confidence interval for the pretest probability between 0.5 and 0.7, and a confidence interval for the likelihood ratio of our test results between 4 and 8, Figure 7.3 tells us that values for the post-test probability between 0.8 and 0.95 are possible.

A third type of sensitivity analysis could be done using the relative seriousness of false positive and false negative results by checking how the threshold between positive and negative test results will shift when different values for this relative seriousness are considered.

"Many" diagnostic tests: logistic regression

The analysis of many diagnostic tests is more complicated than the analysis of a single diagnostic test. There is, however, a standard statistical method, logistic regression, that can be applied in this situation. It is a general method for the analysis of binary data, such as the presence or absence of

disease.[6] It is best seen as a generalised form of Bayes' theorem, using a logarithmic transformation, in order to have an additive instead of a multiplicative formula. Thus "post-test odds equals pretest odds times likelihood ratio", becomes, after taking the logarithm, "log post-test odds equals log pretest odds *plus* log likelihood ratio". Or, in formula form, with the L indicating "logarithm":

$$O(X) = O \times LR \text{ becomes } LO(X) = LO + LLR(X)$$

The similar generalised *formula of logistic regression* is as follows: the log odds of disease, (also called logit (LO)), given test results X_1, X_2, … X_k, is a linear function of the test results:

$$LO(X_1, X_2, \cdots X_k) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

Usually, the natural logarithm (Ln) is used, as we will do in the remainder of this section. We will illustrate logistic regression by applying it to a single dichotomous test, because in such a case calculations are easily done by hand. We take again the renography test for renal artery stenosis, the results of which were depicted in Table 7.2.

We start with the situation prior to performing the test: our best estimate of stenosis is then the prior or pretest probability, based on the observation that we have 100 patients with and 337 patients without stenosis. The logistic formula in this case $LnO = b_0$, and contains only a constant b_0, because no tests have been performed yet. Thus b_0, the Ln pretest odds, which in this case is Ln (100/337) = −1.21.

Next the renography test is performed. The test result is coded as $X_1 = 0$ (normal) or $X_1 = 1$ (abnormal), and the logistic formula is $LnO(X_1) = b_0 + b_1X_1$. We will derive the coefficients b_0 and b_1 by applying the log odds form of Bayes' theorem to both the normal and the abnormal test results. The logistic formula follows immediately from the results.

In case of a normal renogram ($X_1 = 0$) there are 29 patients with and 304 patients without stenosis in Table 7.2. Bayes' theorem tells that the log odds on stenosis for result $X_1 = 0$, $LnO(X_1 = 0) = Ln(29/304) = -2.35$, equals the Ln pretest odds of −1.21 plus the Ln likelihood ratio of a normal renogram, which is −1.14.

In case of an abnormal renogram ($X_1 = 1$) there are 71 patients with and 33 patients without stenosis, and $LnO(X_1 = 1) = Ln(71/33) = 0.77$, being the Ln pretest odds of −1.21 plus the Ln likelihood ratio 1.98 of an abnormal renogram. Combining the two applications of Bayes' theorem, we get:

$$LnO(X_1) = -1.21 - 1.14 \text{ (when } X_1 = 0) + 1.98 \text{ (when } X_1 = 1).$$

This can be simplified to the logistic formula:

$$LnO(X_1) = b_0 + b_1X_1 = -2.35 + 3.12\ X_1.$$

Two remarks can be made: the coefficient b_1 (3.12) is precisely the Ln of the diagnostic odds ratio (22.6) of renography discussed earlier. And b_0 in the logistic formula can no longer be interpreted as a pretest log odds of stenosis, but as the $\mathrm{LnO}(X_1=0)$. This completes the logistic regression analysis of one dichotomous test.

When more than one test is involved the calculations are extensions of the described single test situation, using the multiple logistic regression formula:

$$\mathrm{Ln}(X_1, X_2, \cdots X_k) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

Using this approach, the investigator can account for dependency and interaction between the various tests.[7] However, such multivariate calculations are very troublesome to do by hand. In general, when many tests, including continuous tests, are involved, standard statistical software such as SPSS, SAS or BMDP will have to be used. This software also provides 95% confidence intervals for b_0, b_1, $b_2, \ldots b_k$, and the corresponding odds ratios (e^b).

From multiple logistic regression analysis one can not only learn about the predictive value of a combination of tests, but also what a certain test adds to other tests that have already been performed.

Concluding remarks

In this chapter we have given an overview of the most important performance measures of diagnostic tests, illustrated with a clinical example. Also, the estimation of confidence intervals to account for sampling variability has been explained. Furthermore, decision analytical considerations and sensitivity analysis, as methods to deal with value judgements and uncertainties, have been introduced. Finally, the principles of logistic regression to analyse the predictive value of multiple tests, when applied simultaneously, have been outlined.

In applying the presented analysis techniques it is presupposed that the research data have been collected with the avoidance of important bias (affecting internal validity) and with acceptable generalisability to the target population where the diagnostic test(s) are to be applied (external validity). These issues regarding the validity of the study design are dealt with in Chapters 1–6. As a general rule, in the analysis phase one cannot correct for shortcomings of the validity of the study design, such as bias resulting from an unclear or inappropriate process of selection of study subjects, or from an inadequate reference standard. However, if potential factors that may affect test performance are measured during the study, these can be included as independent covariables in the analysis. An example may be age as a potential effect modifier of the performance of renography. The potential influence of other possible biases can be explored using sensitivity analysis.

In the past decade, new data analytical challenges have resulted from the need to synthesise a number of studies and to analyse the pooled data of those studies (meta-analysis). Also, in such a pooled analysis the usual performance measures of diagnostic tests can be assessed, as is shown in Chapter 8. Finally, although it is always the aim to minimise the number of lost outcomes or not-performed tests, in most studies these will not be totally avoided. Although associated methodological problems are discussed in Chapter 2, there are various options for the analytical approach to such "missing values" on which professional biostatisticians can give advice.

References

1 Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*, 2nd edn. London: BMJ Books, 2000 (includes software).
2 Sackett DL, Haynes PB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown and Co, 1985.
3 Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991;**44**:763–70.
4 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
5 Sox HC Jr, Blatt MA, Higgins MC, Marton KI. *Medical decision making*. Stoneham, MA: Butterworths, 1988.
6 Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley, 1989.
7 Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. *Med Decision Making* 1992;**12**:93–108.

Software

- For the analysis of test results, including logistic regression for single and multiple tests and confidence intervals for the diagnostic odds ratio, standard statistical software such as SPSS, SAS or BMDP may be used. In SPSS and SAS analysis of the area under the ROC curve (plus confidence intervals) can be performed.
- Visual Bayes is a freely available program introducing basic methods for the interpretation and validation of diagnostic tests in an intuitive way. It may be downloaded from http://www.imbi.uni-freiburg.de/medinf.
- Treeage-DATA is a decision analysis program in which diagnostic tests and subsequent treatment decisions can be represented. Good opportunities for sensitivity analysis.

Appendix: Tables for confidence intervals for proportions

Table A.1–A.2 Exact confidence intervals for proportions based on small N (**Table A.1**) or on small n (**Table A.2**).

Table A.3 Half 95% confidence intervals for proportions.

For all tables N = number of observations (denominator), n = number of successes (numerator).

Table A.1 Example: In a group of five patients with renal artery stenosis, three were positive on diagnostic renography. The estimated sensitivity of renography is therefore 3/5, that is, 0.6 with 95% confidence interval (0.15–0.95).

Table A.2 Exact confidence intervals for small n. Because the binomial distribution is symmetric around $p = 0.5$, the table can also be used for small values of N-n: when using N-n instead of n the resulting confidence interval changes to (1 − upper limit, 1 − lower limit).

Example for small n: A new screening test for a disease is required to have a very low false positive rate that is, high specificity. In a sample of 200 proven non-diseased subjects only one had a positive test result. The false positive rate is estimated at 1/200 = 0.005 and the exact 95% confidence interval is (0.000–0.028)

Example for small N-n: In the previous example we can estimate the specificity as 199/200 = 0.995, with 95% confidence interval (0.972–1.000).

Table A.3 The half 95% confidence interval for proportions, based on the normal approximation to the binomial distribution for large numbers of observations. With this approximation, the half 95% confidence interval for a proportion $\hat{p} = n/N$ is $1.96 \times \text{SE}$, with SE being the standard error. The 95% confidence interval is constructed by subtracting (lower confidence limit) or adding (upper confidence limit) the number from the table to the estimate p. By symmetry the values for n and for N-n are the same. When the values of n and N are not given directly in the table, linear interpolation for n and/or N may be used.

Example

In a group of 333 patients with a negative renography, 29 nevertheless appeared to suffer from renal artery stenosis. The estimated negative predictive value (NPV) of renography is therefore (333–29)/333, that is, 0.91.

The value from the table is required for $N = 333$ and N-$n = 29$. We use linear interpolation for N. At $N = 300$, the table gives a value of 0.0335 and at $N = 500$ the value is 0.0205, for $n = 29$, taking the averages of the values for $n = 28$ and 30.

Linear interpolation at $N = 333$ requires:
({Value at $N = 333$} − 0.0355) : (0.0205 − 0.0335) = (333 − 300) : (500 − 300). Thus {Value at $N = 333$} = 0.03.
The 95% confidence interval becomes (0.91 − 0.03, 0.91 + 0.03) = (0.88 − 0.94).

Note 1

Instead of interpolation, the formula for SE could have been used directly:

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{\text{N}}}, \text{ giving SE} = \sqrt{\frac{\frac{304}{333}\left(1-\frac{29}{333}\right)}{333}} = 0.0154$$

Multiplying by 1.96 gives a value of 0.03 for the half 95% confidence interval.

Note 2

For other levels of confidence the numbers in Table A.3 have to be multiplied by a factor. The following table gives multiplication factors for a few commonly used levels of confidence:

Confidence level (%)	Multiplication factor
50	0.34
67	0.49
80	0.65
90	0.84
95	1
99	1.31
99.9	1.68

Table A.1 Exact 95% confidence intervals for proportions n/N for N from 2 to 25

n	p	95% CI		n	p	95% CI		n	p	95% CI	
	N=2				N=8(cont'd)				N=12(cont'd)		
0	0	0.00	0.78	3	0.38	0.09	0.76	2	0.17	0.02	0.48
1	0.5	0.01	0.99	4	0.5	0.16	0.84	3	0.25	0.05	0.57
2	1	0.22	1.00	5	0.63	0.24	0.91	4	0.33	0.10	0.65
				6	0.75	0.35	0.97	5	0.42	0.15	0.72
	[N = 3]			7	0.88	0.47	1.00	6	0.5	0.21	0.79
0	0	0.00	0.63	8	1	0.69	1.00	7	0.58	0.28	0.85
1	0.33	0.01	0.91					8	0.67	0.35	0.90
2	0.67	0.09	0.99		N = 9			9	0.75	0.43	0.95
3	1	0.37	1.00	0	0	0.00	0.28	10	0.83	0.52	0.98
				1	0.11	0.00	0.48	11	0.92	0.62	1.00
	[N = 4]			2	0.22	0.03	0.60	12	1	0.78	1.00
0	0	0.00	0.53	3	0.33	0.07	0.70				
1	0.25	0.01	0.81	4	0.44	0.14	0.79		N = 13		
2	0.5	0.07	0.93	5	0.56	0.21	0.86	0	0	0.00	0.21
3	0.75	0.19	0.99	6	0.67	0.30	0.93	1	0.08	0.00	0.36
4	1	0.47	1.00	7	0.78	0.40	0.97	2	0.15	0.02	0.45
				8	0.89	0.52	1.00	3	0.23	0.05	0.54
	[N = 5]			9	1	0.72	1.00	4	0.31	0.09	0.61
0	0	0.00	0.45					5	0.38	0.14	0.68
1	0.2	0.01	0.72		N = 10			6	0.46	0.19	0.75
2	0.4	0.05	0.85	0	0	0.00	0.26	7	0.54	0.25	0.81
3	0.6	0.15	0.95	1	0.1	0.00	0.45	8	0.62	0.32	0.86
4	0.8	0.28	0.99	2	0.2	0.03	0.56	9	0.69	0.39	0.91
5	1	0.55	1.00	3	0.3	0.07	0.65	10	0.77	0.46	0.95
				4	0.4	0.12	0.74	11	0.85	0.55	0.98
	[N = 6]			5	0.5	0.19	0.81	12	0.92	0.64	1.00
0	0	0.00	0.39	6	0.6	0.26	0.88	13	1	0.79	1.00
1	0.17	0.00	0.64	7	0.7	0.35	0.93				
2	0.33	0.04	0.78	8	0.8	0.44	0.97		N = 14		
3	0.5	0.12	0.88	9	0.9	0.55	1.00	0	0	0.00	0.19
4	0.67	0.22	0.96	10	1	0.74	1.00	1	0.07	0.00	0.34
5	0.83	0.36	1.00					2	0.14	0.02	0.43
6	1	0.61	1.00		N = 11			3	0.21	0.05	0.51
				0	0	0.00	0.24	4	0.29	0.08	0.58
	[N = 7]			1	0.09	0.00	0.41	5	0.36	0.13	0.65
0	0	0.00	0.35	2	0.18	0.02	0.52	6	0.43	0.18	0.71
1	0.14	0.00	0.58	3	0.27	0.06	0.61	7	0.5	0.23	0.77
2	0.29	0.04	0.71	4	0.36	0.11	0.69	8	0.57	0.29	0.82
3	0.43	0.10	0.82	5	0.45	0.17	0.77	9	0.64	0.35	0.87
4	0.57	0.18	0.90	6	0.55	0.23	0.83	10	0.71	0.42	0.92
5	0.71	0.29	0.96	7	0.64	0.31	0.89	11	0.79	0.49	0.95
6	0.86	0.42	1.00	8	0.73	0.39	0.94	12	0.86	0.57	0.98
7	1	0.65	1.00	9	0.82	0.48	0.98	13	0.93	0.66	1.00
				10	0.91	0.59	1.00	14	1	0.81	1.00
	[N = 8]			11	1	0.76	1.00				
0	0	0.00	0.31						N = 15		
1	0.13	0.00	0.53		N = 12			0	0	0.00	0.18
2	0.25	0.03	0.65	0	0	0.00	0.22	1	0.07	0.00	0.32
				1	0.08	0.00	0.38	2	0.13	0.02	0.40
								3	0.2	0.04	0.48

Table A.1 Continued

n	*p*	95% CI	
	N=15(cont'd)		
4	0.27	0.08	0.55
5	0.33	0.12	0.62
6	0.4	0.16	0.68
7	0.47	0.21	0.73
8	0.53	0.27	0.79
9	0.6	0.32	0.84
10	0.67	0.38	0.88
11	0.73	0.45	0.92
12	0.8	0.52	0.96
13	0.87	0.60	0.98
14	0.93	0.68	1.00
15	1	0.82	1.00
[*N* = 16]			
0	0	0.00	0.17
1	0.06	0.00	0.30
2	0.13	0.02	0.38
3	0.19	0.04	0.46
4	0.25	0.07	0.52
5	0.31	0.11	0.59
6	0.38	0.15	0.65
7	0.44	0.20	0.70
8	0.5	0.25	0.75
9	0.56	0.30	0.80
10	0.63	0.35	0.85
11	0.69	0.41	0.89
12	0.75	0.48	0.93
13	0.81	0.54	0.96
14	0.88	0.62	0.98
15	0.94	0.70	1.00
16	1	0.83	1.00
[*N* = 17]			
0	0	0.00	0.16
1	0.06	0.00	0.29
2	0.12	0.01	0.36
3	0.18	0.04	0.43
4	0.24	0.07	0.50
5	0.29	0.10	0.56
6	0.35	0.14	0.62
7	0.41	0.18	0.67
8	0.47	0.23	0.72
9	0.53	0.28	0.77
10	0.59	0.33	0.82
11	0.65	0.38	0.86
12	0.71	0.44	0.90
13	0.76	0.50	0.93
14	0.82	0.57	0.96
15	0.88	0.64	0.99
16	0.94	0.71	1.00
17	1	0.84	1.00

n	*p*	95% CI	
	N=18		
0	0	0.00	0.15
1	0.06	0.00	0.27
2	0.11	0.01	0.35
3	0.17	0.04	0.41
4	0.22	0.06	0.48
5	0.28	0.10	0.53
6	0.33	0.13	0.59
7	0.39	0.17	0.64
8	0.44	0.22	0.69
9	0.5	0.26	0.74
10	0.56	0.31	0.78
11	0.61	0.36	0.83
12	0.67	0.41	0.87
13	0.72	0.47	0.90
14	0.78	0.52	0.94
15	0.83	0.59	0.96
16	0.89	0.65	0.99
17	0.94	0.73	1.00
18	1	0.85	1.00
	N = 19		
0	0	0.00	0.15
1	0.05	0.00	0.26
2	0.11	0.01	0.33
3	0.16	0.03	0.40
4	0.21	0.06	0.46
5	0.26	0.09	0.51
6	0.32	0.13	0.57
7	0.37	0.16	0.62
8	0.42	0.20	0.67
9	0.47	0.24	0.71
10	0.53	0.29	0.76
11	0.58	0.33	0.80
12	0.63	0.38	0.84
13	0.68	0.43	0.87
14	0.74	0.49	0.91
15	0.79	0.54	0.94
16	0.84	0.60	0.97
17	0.89	0.67	0.99
18	0.95	0.74	1.00
19	1	0.85	1.00
	N = 20		
0	0	0.00	0.14
1	0.05	0.00	0.25
2	0.1	0.01	0.32
3	0.15	0.03	0.38
4	0.2	0.06	0.44
5	0.25	0.09	0.49
6	0.3	0.12	0.54

n	*p*	95% CI	
	N=20(cont'd)		
7	0.35	0.15	0.59
8	0.4	0.19	0.64
9	0.45	0.23	0.68
10	0.5	0.27	0.73
11	0.55	0.32	0.77
12	0.6	0.36	0.81
13	0.65	0.41	0.85
14	0.7	0.46	0.88
15	0.75	0.51	0.91
16	0.8	0.56	0.94
17	0.85	0.62	0.97
18	0.9	0.68	0.99
19	0.95	0.75	1.00
20	1	0.86	1.00
	N = 21		
0	0	0.00	0.13
1	0.05	0.00	0.24
2	0.1	0.01	0.30
3	0.14	0.03	0.36
4	0.19	0.05	0.42
5	0.24	0.08	0.47
6	0.29	0.11	0.52
7	0.33	0.15	0.57
8	0.38	0.18	0.62
9	0.43	0.22	0.66
10	0.48	0.26	0.70
11	0.52	0.30	0.74
12	0.57	0.34	0.78
13	0.62	0.38	0.82
14	0.67	0.43	0.85
15	0.71	0.48	0.89
16	0.76	0.53	0.92
17	0.81	0.58	0.95
18	0.86	0.64	0.97
19	0.9	0.70	0.99
20	0.95	0.76	1.00
21	1	0.87	1.00
	N = 22		
0	0	0.00	0.13
1	0.05	0.00	0.23
2	0.09	0.01	0.29
3	0.14	0.03	0.35
4	0.18	0.05	0.40
5	0.23	0.08	0.45
6	0.27	0.11	0.50
7	0.32	0.14	0.55
8	0.36	0.17	0.59
9	0.41	0.21	0.64
10	0.45	0.24	0.68

Table A.1 Continued

n	*p*	95% CI		*n*	*p*	95% CI		*n*	*p*	95% CI	
N=22(cont'd)				*N*=23(cont'd)				*N*=24(cont'd)			
11	0.5	0.28	0.72	17	0.74	0.52	0.90	22	0.92	0.73	0.99
12	0.55	0.32	0.76	18	0.78	0.56	0.93	23	0.96	0.79	1.00
13	0.59	0.36	0.79	19	0.83	0.61	0.95	24	1	0.88	1.00
14	0.64	0.41	0.83	20	0.87	0.66	0.97				
15	0.68	0.45	0.86	21	0.91	0.72	0.99				
16	0.73	0.50	0.89	22	0.96	0.78	1.00	*N* = 25			
17	0.77	0.55	0.92	23	1	0.88	1.00	0	0	0.00	0.11
18	0.82	0.60	0.95					1	0.04	0.00	0.20
19	0.86	0.65	0.97					2	0.08	0.01	0.26
20	0.91	0.71	0.99	*N* = 24				3	0.12	0.03	0.31
21	0.95	0.77	1.00	0	0	0.00	0.12	4	0.16	0.05	0.36
22	1	0.87	1.00	1	0.04	0.00	0.21	5	0.2	0.07	0.41
				2	0.08	0.01	0.27	6	0.24	0.09	0.45
				3	0.13	0.03	0.32	7	0.28	0.12	0.49
[*N* = 23]				4	0.17	0.05	0.37	8	0.32	0.15	0.54
0	0	0.00	0.12	5	0.21	0.07	0.42	9	0.36	0.18	0.57
1	0.04	0.00	0.22	6	0.25	0.10	0.47	10	0.4	0.21	0.61
2	0.09	0.01	0.28	7	0.29	0.13	0.51	11	0.44	0.24	0.65
3	0.13	0.03	0.34	8	0.33	0.16	0.55	12	0.48	0.28	0.69
4	0.17	0.05	0.39	9	0.38	0.19	0.59	13	0.52	0.31	0.72
5	0.22	0.07	0.44	10	0.42	0.22	0.63	14	0.56	0.35	0.76
6	0.26	0.10	0.48	11	0.46	0.26	0.67	15	0.6	0.39	0.79
7	0.3	0.13	0.53	12	0.5	0.29	0.71	16	0.64	0.43	0.82
8	0.35	0.16	0.57	13	0.54	0.33	0.74	17	0.68	0.46	0.85
9	0.39	0.20	0.61	14	0.58	0.37	0.78	18	0.72	0.51	0.88
10	0.43	0.23	0.66	15	0.63	0.41	0.81	19	0.76	0.55	0.91
11	0.48	0.27	0.69	16	0.67	0.45	0.84	20	0.8	0.59	0.93
12	0.52	0.31	0.73	17	0.71	0.49	0.87	21	0.84	0.64	0.95
13	0.57	0.34	0.77	18	0.75	0.53	0.90	22	0.88	0.69	0.97
14	0.61	0.39	0.80	19	0.79	0.58	0.93	23	0.92	0.74	0.99
15	0.65	0.43	0.84	20	0.83	0.63	0.95	24	0.96	0.80	1.00
16	0.7	0.47	0.87	21	0.88	0.68	0.97	25	1	0.89	1.00

Table A.2 Exact 95% confidence intervals for proportions n/N for small n (0–7) and N = 30, 35, ... ,70, 80, 90, 100, 120, 150, 200, 300, 500, 1000, 2000.

	n															
	0		1		2		3		4		5		6		7	
N																
30	0.000	0.095	0.001	0.172	0.008	0.221	0.021	0.265	0.038	0.307	0.056	0.347	0.077	0.386	0.099	0.423
35	0.000	0.082	0.001	0.149	0.007	0.192	0.018	0.231	0.032	0.267	0.048	0.303	0.066	0.336	0.084	0.369
40	0.000	0.072	0.001	0.132	0.006	0.169	0.016	0.204	0.028	0.237	0.042	0.268	0.057	0.298	0.073	0.328
45	0.000	0.064	0.001	0.118	0.005	0.151	0.014	0.183	0.025	0.212	0.037	0.241	0.051	0.268	0.065	0.295
50	0.000	0.058	0.001	0.106	0.005	0.137	0.013	0.165	0.022	0.192	0.033	0.218	0.045	0.243	0.058	0.267
55	0.000	0.053	0.000	0.097	0.004	0.125	0.011	0.151	0.020	0.176	0.030	0.200	0.041	0.222	0.053	0.245
60	0.000	0.049	0.000	0.089	0.004	0.115	0.010	0.139	0.018	0.162	0.028	0.184	0.038	0.205	0.048	0.226
65	0.000	0.045	0.000	0.083	0.004	0.107	0.010	0.129	0.017	0.150	0.025	0.170	0.035	0.190	0.044	0.209
70	0.000	0.042	0.000	0.077	0.003	0.099	0.009	0.120	0.016	0.140	0.024	0.159	0.032	0.177	0.041	0.195
80	0.000	0.037	0.000	0.068	0.003	0.087	0.008	0.106	0.014	0.123	0.021	0.140	0.028	0.156	0.036	0.172
90	0.000	0.033	0.000	0.060	0.003	0.078	0.007	0.094	0.012	0.110	0.018	0.125	0.025	0.139	0.032	0.154
100	0.000	0.030	0.000	0.054	0.002	0.070	0.006	0.085	0.011	0.099	0.016	0.113	0.022	0.126	0.029	0.139
120	0.000	0.025	0.000	0.046	0.002	0.059	0.005	0.071	0.009	0.083	0.014	0.095	0.019	0.106	0.024	0.116
150	0.000	0.020	0.000	0.037	0.002	0.047	0.004	0.057	0.007	0.067	0.011	0.076	0.015	0.085	0.019	0.094
200	0.000	0.015	0.000	0.028	0.001	0.036	0.003	0.043	0.005	0.050	0.008	0.057	0.011	0.064	0.014	0.071
300	0.000	0.010	0.000	0.018	0.001	0.024	0.002	0.029	0.004	0.034	0.005	0.038	0.007	0.043	0.009	0.047
500	0.000	0.006	0.000	0.011	0.000	0.014	0.001	0.017	0.002	0.020	0.003	0.023	0.004	0.026	0.006	0.029
1000	0.000	0.003	0.000	0.006	0.000	0.007	0.001	0.009	0.001	0.010	0.002	0.012	0.002	0.013	0.003	0.014
2000	0.000	0.001	0.000	0.003	0.000	0.004	0.000	0.004	0.001	0.005	0.001	0.006	0.001	0.007	0.001	0.007

Table A.3 Half 95% confidence intervals (=1.96 × SE) of proportions n/N for N = 30,35,..., 70, 80, 90, 100, 120, 150, 200, 300, 500, 1000, 2000.

	N																		
	30	35	40	45	50	55	60	65	70	80	90	100	120	150	200	300	500	1000	2000
$N-n$																			
8	0.158	0.139	0.124	0.112	0.102	0.093	0.086	0.080	0.075	0.066	0.059	0.053	0.045	0.036	0.027	0.018	0.011	0.006	0.003
9	0.164	0.145	0.129	0.117	0.106	0.098	0.090	0.084	0.078	0.069	0.062	0.056	0.047	0.038	0.029	0.019	0.012	0.006	0.003
10	0.169	0.150	0.134	0.121	0.111	0.102	0.094	0.088	0.082	0.072	0.065	0.059	0.049	0.040	0.030	0.020	0.012	0.006	0.003
11	0.172	0.154	0.138	0.126	0.115	0.106	0.098	0.091	0.085	0.075	0.068	0.061	0.052	0.042	0.032	0.021	0.013	0.006	0.003
12	0.175	0.157	0.142	0.129	0.118	0.109	0.101	0.094	0.088	0.078	0.070	0.064	0.054	0.043	0.033	0.022	0.013	0.007	0.003
13	0.177	0.160	0.145	0.132	0.122	0.112	0.104	0.097	0.091	0.081	0.073	0.066	0.056	0.045	0.034	0.023	0.014	0.007	0.004
14	0.179	0.162	0.148	0.135	0.124	0.115	0.107	0.100	0.094	0.083	0.075	0.068	0.057	0.047	0.035	0.024	0.014	0.007	0.004
15	0.179	0.164	0.150	0.138	0.127	0.118	0.110	0.102	0.096	0.086	0.077	0.070	0.059	0.048	0.037	0.025	0.015	0.008	0.004
16		0.165	0.152	0.140	0.129	0.120	0.112	0.105	0.098	0.088	0.079	0.072	0.061	0.049	0.038	0.025	0.015	0.008	0.004
17		0.166	0.153	0.142	0.131	0.122	0.114	0.107	0.100	0.090	0.081	0.074	0.062	0.051	0.039	0.026	0.016	0.008	0.004
18		0.166	0.154	0.143	0.133	0.124	0.116	0.109	0.102	0.092	0.083	0.075	0.064	0.052	0.040	0.027	0.016	0.008	0.004
19			0.155	0.144	0.135	0.126	0.118	0.111	0.104	0.093	0.084	0.077	0.065	0.053	0.041	0.028	0.017	0.008	0.004
20			0.155	0.145	0.136	0.127	0.119	0.112	0.106	0.095	0.086	0.078	0.067	0.054	0.042	0.028	0.017	0.009	0.004
22				0.146	0.138	0.129	0.122	0.115	0.109	0.098	0.089	0.081	0.069	0.057	0.043	0.029	0.018	0.009	0.005
24					0.138	0.131	0.124	0.117	0.111	0.100	0.091	0.084	0.072	0.059	0.045	0.031	0.019	0.009	0.005
26						0.132	0.125	0.119	0.113	0.103	0.094	0.086	0.074	0.061	0.047	0.032	0.019	0.010	0.005
28							0.126	0.120	0.115	0.105	0.096	0.088	0.076	0.062	0.048	0.033	0.020	0.010	0.005
30							0.127	0.121	0.116	0.106	0.097	0.090	0.077	0.064	0.049	0.034	0.021	0.011	0.005
32								0.122	0.117	0.107	0.099	0.091	0.079	0.066	0.051	0.035	0.021	0.011	0.005
34									0.117	0.108	0.100	0.093	0.081	0.067	0.052	0.036	0.022	0.011	0.006
36										0.109	0.101	0.094	0.082	0.068	0.053	0.037	0.023	0.012	0.006

38	0.109	0.102	0.095	0.083	0.070	0.054	0.038	0.023	0.012	0.006
40	0.110	0.103	0.096	0.084	0.071	0.055	0.038	0.024	0.012	0.006
45		0.103	0.098	0.087	0.073	0.058	0.040	0.025	0.013	0.006
50			0.098	0.088	0.075	0.060	0.042	0.026	0.014	0.007
55				0.089	0.077	0.062	0.044	0.027	0.014	0.007
60				0.089	0.078	0.064	0.045	0.028	0.015	0.007
70					0.080	0.066	0.048	0.030	0.016	0.008
80						0.068	0.050	0.032	0.017	0.009
90						0.069	0.052	0.034	0.018	0.009
100						0.069	0.053	0.035	0.019	0.010
120							0.055	0.037	0.020	0.010
140							0.056	0.039	0.022	0.011
160							0.056	0.041	0.023	0.012
180								0.042	0.024	0.013
200								0.043	0.025	0.013
300									0.028	0.016
400									0.030	0.018
500									0.031	0.019
1000										0.022

8 Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests

WALTER L DEVILLÉ, FRANK BUNTINX
On behalf of an ad hoc working group of the Cochrane Methods Group on Screening and Diagnostic Tests*

Summary box

- A systematic review should include all available evidence, and so a systematic and comprehensive search of the literature is needed in computerised databases and other sources.
- The search strategy must be based on an explicit description of the subjects receiving the test of interest, the diagnostic test and its accuracy estimates, the target disease, and the study design. These elements can be specified in the criteria for inclusion of primary studies in the review.
- Two independent reviewers should screen the titles and abstracts of the identified citations using specific prespecified inclusion criteria. In case of disagreement or insufficient information, a third reviewer and/or the full papers should be consulted. The publications to be evaluated should provide sufficient information on the reference standard, the study population, and the setting(s) studied.

*Riekie de Vet, Jeroen Lijmer, Victor Montori

- The methodological quality of each selected paper should be assessed independently by at least two reviewers. Chance-adjusted agreement should be reported and disagreement solved by consensus or arbitration. Internal and external validity criteria, describing participants, diagnostic test, and target disease of interest, and study methods can be used in meta-analysis to assess the overall "level of evidence", and in sensitivity and subgroup analyses.
- Two reviewers should independently extract the required information from the primary studies, about the participants, the testing procedure, the cut-off points used, exclusions, and indeterminate results.
- To be able to carry out subgroup analyses, sources of heterogeneity should be defined based on *a priori* existing hypotheses.
- Whether meta-analysis with statistical pooling can be conducted depends on the number and methodological quality of the primary studies.
- In view of the low methodological quality of most published diagnostic studies, the use of random effect models for pooling may be useful, even if there is no apparent heterogeneity.
- Methods for statistical pooling of proportions, likelihood ratios, and ROC curves are provided.

Introduction

Systematic reviews and meta-analyses of studies evaluating the accuracy of diagnostic tests (we will refer to them generically as diagnostic systematic reviews) are appearing more often in the medical literature.[1,2] Of the 26 reviews on diagnostic tests published between 1996 and 1997, 19 were systematic reviews or meta-analyses.[2] In the field of clinical chemistry and haematology, 23 of 45 reviews published between 1985 and 1998 were systematic reviews.[3] Although guidelines for the critical appraisal of diagnostic research and meta-analyses have already been published,[1,4–7] these may be difficult for clinical researchers to understand.

We here present a set of practical guidelines, based on evidence and the expertise of the Cochrane Collaboration, to facilitate the understanding of and appropriate adherence to methodological principles when conducting diagnostic systematic reviews. We reviewed reports of systematic searches of the literature for diagnostic research,[8–11] methodological criteria to evaluate diagnostic research,[1,4–7] methods for statistical pooling of data on diagnostic accuracy,[12–20] and methods for exploring heterogeneity.[21–25]

Guidelines for conducting diagnostic systematic reviews are presented in a stepwise fashion and are followed by comments providing further information. Examples are given using the results of two systematic reviews on the accuracy of the urine dipstick in the diagnosis of urinary tract infections,[26] and on the accuracy of the straight-leg raising test in the diagnosis of intervertebral disc hernia.[27]

The guidelines

How to search the literature for studies evaluating the accuracy of diagnostic tests

A systematic review should include all available evidence, and so a systematic and comprehensive search of the literature is needed. The reviewer has to design a search strategy based on a clear and explicit description of the subjects receiving the test of interest, the diagnostic test and its accuracy estimates, the target disease, and the study design. These elements are usually specified in the criteria for inclusion of primary studies in the review. The search will include electronic literature databases. However, because computerised databases only index a subset of all the available literature, the search should be extended using other sources.[9] The search to identify primary studies may take the following basic but labour intensive steps:

1. A computer aided search of MEDLINE (PubMed website (http://www.ncbi.nlm.nih.gov/PUBMED), EMBASE and other databases. A search strategy begins by creating a list of database specific keywords and text words that describe the diagnostic test and the target disease of interest (subject specific strategy). Because the number of diagnostic accuracy studies is often small, the subject specific strategy usually yields a limited number of publications to be screened.[10] An accurate search strategy for diagnostic publications (generic strategy) was recently published[11] and can be combined with the subject specific strategy if the number of publications resulting from the latter is large. We found a combination of two published generic strategies adapted for use in PubMed (MEDLINE) to be more sensitive and precise than previously published strategies[8,10] (Box 8.1). Each electronic database will need to be searched using a specially designed search strategy.
2. The reference section of primary studies, narrative reviews, and systematic reviews should be checked to search for additional primary studies that could have been missed by the electronic search. Identification methods for systematic reviews have also been published.[28] The MEDION database, available at the University of

Maastricht, the Netherlands, collects some 250 published reviews of diagnostic and screening studies. It is available through berna.schouten@hag.unimaas.nl and will shortly be published on the internet.

3. Consultation of experts in the disease of interest to identify further published and unpublished primary studies. As diagnostic accuracy studies are often based on routinely collected data, publication bias may be more prevalent in diagnostic than in therapeutic research.[17]

Box 8.1 Search strategy in PubMed (MEDLINE) for publications about the evaluation of diagnostic accuracy

((((((((((("sensitivity and specificity"[All Fields] OR "sensitivity and specificity/standards"[All Fields]) OR "specificity"[All Fields]) OR "screening"[All Fields]) OR "false positive"[All Fields]) OR "false negative"[All Fields]) OR "accuracy"[All Fields]) OR ((((("predictive value"[All Fields] OR "predictive value of tests"[All Fields]) OR "predictive value of tests/standards"[All Fields]) OR "predictive values"[All Fields]) OR "predictive values of tests"[All Fields])) OR (("reference value"[All Fields] OR "reference values"[All Fields]) OR "reference values/standards"[All Fields])) OR ((((((((((("roc"[All Fields] OR "roc analyses"[All Fields]) OR "roc analysis"[All Fields]) OR "roc and"[All Fields]) OR "roc area"[All Fields]) OR "roc auc"[All Fields]) OR "roc characteristics"[All Fields]) OR "roc curve"[All Fields]) OR "roc curve method"[All Fields]) OR "roc curves"[All Fields]) OR "roc estimated"[All Fields]) OR "roc evaluation"[All Fields])) OR "likelihood ratio"[All Fields]) AND notpubref[sb]) AND "human"[MeSH Terms])

Comments

The first step in a literature search is the identification of relevant publications. Diagnostic research reports – older publications in particular – are often poorly indexed in the electronic databases. It is often fruitful to conduct pilot searches using the subject specific strategy. This process is repeated after identifying and incorporating additional keywords and text words to describe and index the retrieved reports. Studies found only in the reference sections of the retrieved reports but missed by the search strategy should be searched for in the database, using the article title or the first author's name. If a study is found in the database, its keywords should be noted and added to the strategy. Citation tracking may provide additional

studies. The Science Citation Index could be searched forward in time to identify articles citing relevant publications.[29] Once the search is completed, two independent reviewers should screen the titles and abstracts of the identified citations using specific prespecified inclusion criteria. These can be pilot tested on a sample of articles. If disagreements cannot be resolved by consensus, or if insufficient information is available, a third reviewer and/or the full papers should be consulted.

Inclusion criteria

- *Reference test* The accuracy of a diagnostic or screening test should be evaluated by comparing its results with a "gold standard", criterion standard, or reference test accepted as the best available by content experts. The reference test may be a single test, a combination of different tests, or the clinical follow up of patients.[20] The publication should describe the reference test, as it is an essential prerequisite for the evaluation of a diagnostic test.
- *Population* Detailed information about the participants in diagnostic research is often lacking. Participants should be defined explicitly in terms of age, gender, complaints, signs, and symptoms, and their duration. At least a definition of participants with and without the disease, as determined by the reference test, should be available. The minimal number of participants needed with and without the disease depends on the type of study, the estimates of diagnostic accuracy, and the precision used to estimate these parameters.[30]
- *Outcome data* Information should be available to allow the construction of the diagnostic 2×2 table with its four cells: true positives, false negatives, false positives and true negatives.
- *Language* If a review is limited to publications in certain languages, this should be reported.

Comments

As the patient mix (spectrum of disease severity) is different at different levels of care, a diagnostic review may focus on a specific setting (primary care, etc.) or include all levels. This information may be important for subgroup analyses in case of heterogeneity. All evidence available should be reviewed, regardless of the language of publication. It is not easy to identify non-English publications, as they are often not indexed in computerised databases. In the field of intervention research there is some evidence of bias when excluding non-English publications.[31] Our research on the accuracy of the urine dipstick revealed differences in methodological validity between European and American studies, but these differences had no effect on accuracy.

Although large samples are no guarantee against selection bias, small samples seldom result from a consecutive series of patients or a random sample. Small studies are very vulnerable to selection bias.

Methodological quality

The methodological quality of each selected paper should be assessed independently by at least two reviewers. Chance-adjusted agreement should be reported, and disagreements solved by consensus or arbitration. To improve agreement, reviewers should pilot their quality assessment tools in a subset of included studies or studies evaluating a different diagnostic test.

Validity criteria for diagnostic research have been published by the Cochrane Methods Group on Screening and Diagnostic Tests[32] (http://www.som.fmc.flinders.edu.au/fusa/cochrane), and by other authors.[4–6] Criteria assessing internal and external validity should be coded and described explicitly in the review (Table 8.1). The internal validity criteria refer to study characteristics that safeguard against the intrusion of systematic error or bias. External validity criteria provide insight into the generalisability of the study and judge whether the test under evaluation was performed according to accepted standards. Internal and external validity criteria, describing participants, diagnostic test and target disease of interest, and study methods may be used in meta-analysis to assess the overall "level of evidence" and in sensitivity and subgroup analyses (see Data extraction and Data analysis sections).

It is important to remember that studies may appear to be of poor methodological quality because they were either poorly conducted or poorly reported. Methodological appraisal of the primary studies is frequently hindered by lack of information. In these instances reviewers may choose to contact the studies' authors, or to score items as "don't know" or "unclear".

Example A urine dipstick is usually read before the material is cultured. So, it can be interpreted that the dipstick was read without awareness of the results of the culture. However, the culture (reference test) may be interpreted with full awareness of the results of the dipstick. If blinding is not explicitly mentioned, reviewers may choose to score this item as "don't know" or "diagnostic test blinded for reference test" (implicitly scoring the reference test as not blinded). Or, the authors may be contacted for clarification.

A survey of the diagnostic literature from 1990 to 1993 in a number of peer-reviewed journals showed that only a minority of the studies satisfied methodological standards.[7] There is some evidence that inadequate methods may have an impact on the reported accuracy of a diagnostic test: Lijmer[2] screened diagnostic meta-analyses published in 1996 and 1997, and showed that the diagnostic accuracy of a test was overestimated in studies (1) with a case–control design; (2) using different reference tests for

Table 8.1 Validity criteria operationalised for papers reporting on the accuracy of urine dipsticks in the diagnosis of urinary tract infections (UTI) or bacteriuria.

		Positive score
Criteria of internal validity (IV)		
1	Valid reference standard	(Semi-)quantitative (2 points) to dipslide culture (1 point)
2	Definition of cut-off point for reference standard	Definition of urinary tract infection/ bacteriuria by colony forming units per ml (1 point)
3	Blind measurement of index test and reference test	In both directions (2 points) or only index or reference test
4	Avoidance of verification bias	Assessment by reference standard independent from index test results (1 point)
5	Index test interpreted independently of all clinical information	Explicitly mentioned in the publication, or urine samples from mixed outpatient populations examined in a general laboratory (1 point)
6	Design	Prospective (consecutive series) (1 point) or retrospective collection of data (0 points)
Criteria of external validity (EV)		
1	Spectrum of disease	In- and/or exclusion criteria mentioned (1 point)
2	Setting	Enough information to identify setting (1 point)(community through tertiary care)
3	Previous tests/referral filter	Details given about clinical and other diagnostic information as to which index test is being evaluated (symptomatic or asymptomatic patients (1 point)
4	Duration of illness before diagnosis	Duration mentioned (1 point)
5	Comorbid conditions	Details given (type of population) (1 point)
6	Demographic information	Age (1 point) and/or gender (1 point) data provided
7	Execution of index test	Information about standard procedure directly or indirectly available, urine collection procedure, first voided urine, distribution of microorganisms, procedure of contaminated urine samples, time of transportation of urine sample, way of reading index test, persons reading index test (1 point each)
8	Explanation of cut-off point of index test	Trace, 2 or more + (1 point if applicable)
9	Percentage missing	If appropriate: missings mentioned (1 point)
10	Reproducibility of index test	Reproducibility studied or reference mentioned (1 point)

Blinding (IV3): When information about *blinding* of measurements was not given and the dipstick was performed in setting other than the culture, we assumed blind assessment of the index test versus the reference test, but not vice versa.
Explanation of the cut-off point (EV8) was only necessary for the leukocyte esterase measurement.

positive and negative results of the index test; (3) accepting the results of observers who were unblinded to the index test results when performing the reference test; (4) that did not describe diagnostic criteria for the index test; and (5) where participants were inadequately described.

Comments

Ideally, all participants should be submitted to the same reference test. Sometimes different groups of patients are submitted to different reference tests, but details are not given. In this case it is important to assess whether the different reference tests are recognised by experts as being adequate. Verification or work-up bias may be present if not all participants who received the index test are referred to the reference test(s). Verification bias is present if the participants are referred according to the index test results. This is usually the case in screening studies where only subjects with positive index test results receive the reference test, so that only a positive predictive value can be calculated. Estimation of accuracy will not be possible in these studies unless complete follow up registries are available. This is the case if, for example, cancer screening registries and cancer diagnosis registries are coupled.

Data extraction

Two reviewers should independently extract the required information from the primary studies. Detailed information must be extracted about the participants included in the study and about the testing procedures. The cut-off point used in dichotomous testing, and the reasons and the number of participants excluded because of indeterminate results or infeasibility, are always required.

Example Detailed information extracted in the case of the dipstick meta-analysis: mean age, male/female ratio, different cut-off points for leukocyte esterase (trace, 2+, 3+), time needed for transportation, whether indeterminate results were excluded, included as negative, or repeated.

As the information extracted may be used in subgroup analyses and statistical pooling of the validity, possible sources of heterogeneity should be defined based on existing evidence or hypotheses.

Example In the dipstick meta-analysis we hypothesised that the following factors may explain heterogeneity if present: procedures of collection of test material (method of urine collection, delay between urine collection and culture), who was executing the test and how (manually or automatic), and different brands of commercial products.

Accuracy may be presented in different ways. For the meta-analysis of dichotomous tests (see below) it is necessary to construct the diagnostic

2×2 table: absolute numbers in the four cells are needed. Totals of "diseased" and "non-diseased" participants are needed to calculate prior probability (pretest probability), and to reconstruct the 2×2 table from sensitivity, specificity, likelihood ratios, predictive values or receiver operator characteristic (ROC) curves. If possible, the 2×2 table should be generated for all relevant subgroups. Further information to extract includes year of publication, language of publication, and country or region of the world where the study was performed.

Comments

A standardised data extraction form may be used simultaneously with but separately from the quality assessment form. This approach facilitates data extraction and comparison between reviewers. The form has to be piloted to ensure that all reviewers interpret data in the same way. As in other steps of the review where judgements are made, disagreements should be recorded and resolved by consensus or arbitration. Lack of details about test results or cut-off points, inconsequential rounding off of percentages, and data errors require common sense and careful data handling when reconstructing 2×2 tables. If predictive values are presented with sensitivity and specificity in "diseased" and "non-diseased" individuals, the calculation of the four cells from sensitivity and specificity can be confirmed by using the predictive values. Details can be requested from the authors of the studies, but these attempts are often unsuccessful, as the raw data may no longer be available.

Example In a review of the accuracy of the CAGE questionnaire for the diagnosis of alcohol abuse, sufficient data were made available in only nine of the 22 studies selected, although the authors of the review tried to contact the original authors by all means.[33]

Data analysis

Whether or not a meta-analysis – statistical analysis and calculation of a summary diagnostic accuracy estimate – can be conducted depends on the number and methodological quality of primary studies included and the degree of heterogeneity of their estimates of diagnostic accuracy. Because diagnostic accuracy studies are often heterogeneous and present limited information it is typically difficult to complete a meta-analysis. If heterogeneity is identified, important information is obtained from attempts to explain it. For instance, the effect that each validity criterion has on the estimates of diagnostic accuracy and the influence of previously defined study characteristics should be explored as potential explanations of the observed study to study variation.[21–25] If meta-analysis is not possible

or advisable, the review can be limited to a qualitative descriptive analysis of the diagnostic research available (best evidence synthesis).[34]

Several meta-analytical methods for diagnostic research have been published in the last decade.[12–20] For the analysis we recommend the following steps: (1) presentation of the results of individual studies; (2) searching for the presence of heterogeneity; (3) testing for the presence of an (implicit) cut-point effect; (4) dealing with heterogeneity; (5) deciding which model should be used if statistical pooling is appropriate; and (6) statistical pooling.

Describing the results of individual studies

Reporting the main results of all included studies is an essential part of each review. It provides the reader with the outcome measures and gives an insight into their heterogeneity. Each study is presented with some background information (year of publication, geographical region, number of diseased and non-diseased patients, selection of the patients, methodological characteristics) and a summary of the results. In view of the asymmetrical nature of most diagnostic tests (some tests are good to exclude a disease, others to confirm it), it is important to report pairs of complementary outcome measures, that is, both sensitivity and specificity, positive and negative predictive value, likelihood ratio of a positive and of a negative test, or a combination of these. The diagnostic odds ratio (DOR) can be added, but better not alone, as a same odds ratio can relate to different combinations of sensitivity and specificity. Main outcome measures should be reported with their 95% confidence intervals (CI).

$$\mathrm{DOR} = \frac{\text{sensitivity}/(1-\text{sensitivity})}{(1-\text{specificity})/\text{specificity}}$$

Searching for heterogeneity

Basically, heterogeneity relates to the input characteristics of each study (study population, test methods, etc.). When setting inclusion criteria, most reviewers will try to define a more or less homogeneous set of studies. The reality, however, is that even then most diagnostic reviews suffer from considerable heterogeneity. When different studies have largely different results, this can be because of either random error or heterogeneity. To test for homogeneity of sensitivity and specificity, a χ^2 test or an extension of Fisher's exact test for small studies[35] can be used. This may offer some guidance, although the power of this test tends to be low. A basic but very informative method when searching for heterogeneity is to produce a graph in which the individual study outcomes are plotted, together with their 95% confidence intervals (Figure 8.1).

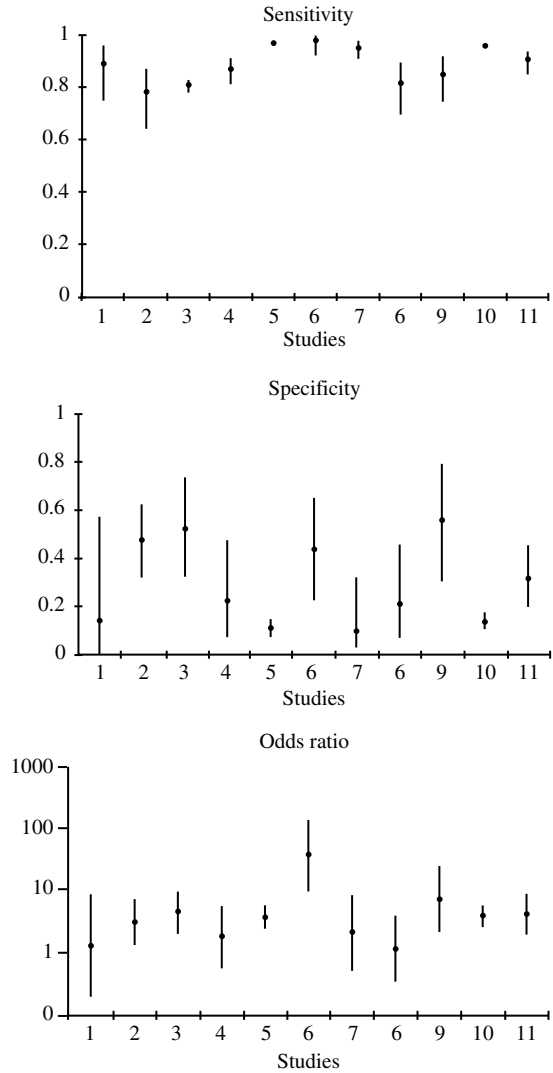


Figure 8.1 Point estimates (with confidence limits) of, respectively, sensitivity, specificity, and diagnostic odds ratio of 11 studies on the validity of the test of Lasègue for the diagnosis of disc hernia in low back pain. Study 6 is an outlier.

Searching for the presence of an (implicit) cut-off point effect

Estimates of diagnostic accuracy differ if not all studies use the same cut-off point for a positive test result or for the reference standard. The interpretation of test results often depends on human factors (for example radiology, pathology, etc.) or on the process of testing (for example clinical

examination). In such cases different studies may use a different implicit cut-off point. Variation in the parameters of accuracy may be partly due to variation in cut-off points. One can test for the presence of a cut-off point effect between studies by calculating a Spearman correlation coefficient between the sensitivity and the specificity of all included studies. If strongly negatively correlated ($\rho > -0.4$), pairs of parameters represent the same DOR. A strong correlation between both parameters will usually result in a homogeneous logarithmic transformed DOR (lnDOR). The test for homogeneity of the lnDOR is described by Fleiss.[35] If the LnDORs of the included studies are homogeneous, a summary ROC curve can be fitted based on the pairs of sensitivity and specificity of the individual studies (see page 159). If sufficient information is available, the pooling of ROC curves of individual studies will also be possible.

Dealing with heterogeneity

In many cases the interpretation of present heterogeneity is the most fascinating and productive part of a meta-analysis. The inspection of the plot of all outcome parameters with their 95% CI may indicate the presence of outliers. In such cases the reason for this situation should be carefully examined.

Example In a review of the diagnostic value of macroscopic haematuria for the diagnosis of urological cancers in primary care, the positive predictive values (PPV) indicated a homogeneous series of five studies with a pooled PPV of 0.19 (95% CI = 0.17–0.23) and one other with a PPV of 0.40.[36] The reason for this high PPV was mentioned in the original study: "GPs' services in the region are extremely good and cases of less serious conditions are probably adequately shifted out and treated without referral", leading to a highly selected study population with a high prior probability.

In such cases an outlier can be excluded and the analysis continued with the homogeneous group of remaining studies. Deviant results should be explored and explained. The decision to exclude outliers is complex and should be handled in the same way as in other fields of research.

Outliers can also be searched by using a Galbraith plot.[37] To construct this plot, the standardised lnDOR = lnDOR/se is plotted (y axis) against the inverse of the se (1/se) (x axis). A regression line that goes through the origin is calculated, together with 95% boundaries (starting at +2 and −2 on the y axis). Studies outside these 95% boundaries may be considered as outliers (Figure 8.2).

Subgroup analyses defined in the protocol could be conducted to detect homogeneous subgroups. Analysis of variance, with the lnDOR as a dependent variable and categorical variables for subgroups as factors, can be used to look for differences among subgroups.

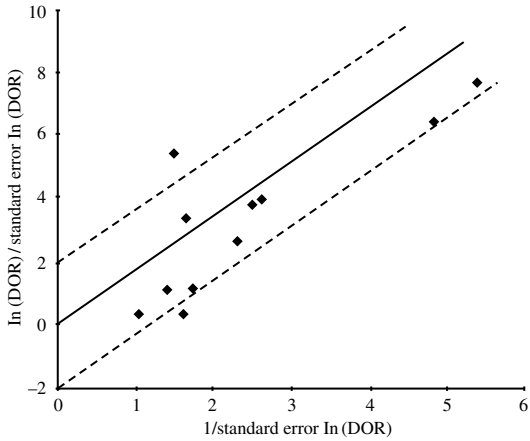


Figure 8.2 Galbraith plot of 11 studies on the validity of the Lasègue test for the diagnosis of disc hernia in low back pain. Study 6 is an outlier.

Example In the dipstick review, sensitivity and specificity were weakly associated ($\rho = -0.227$) and very heterogeneous. Subgroup analysis showed significant differences of the lnDOR between six different populations of participants. In three populations there was a strong negative association between sensitivity and specificity ($\rho = -0.539$, -0.559, and -1.00, respectively), yielding homogeneous lnDOR in the three subgroups. Different SROC curves for each subgroup could be fitted (see section on statistical pooling) (Figure 8.3).

If many studies are available, a more complex multivariate model can be built in which a number of study characteristics are entered as possible covariates. Multivariate models search for the independent effect of study characteristics, adjusted for the influence of other, more powerful ones.

Deciding on the model to be used for statistical pooling

Models

There are two underlying models that can be used when pooling the results of individual studies.

A *fixed effect model* assumes that all studies are a certain random sample of one large common study, and that differences between study outcomes only result from random error. Pooling is simple. It consists essentially of calculating a weighted average of the individual study results. Studies are weighted by the inverse of the variance of the parameter of test accuracy, or by the number of participants.

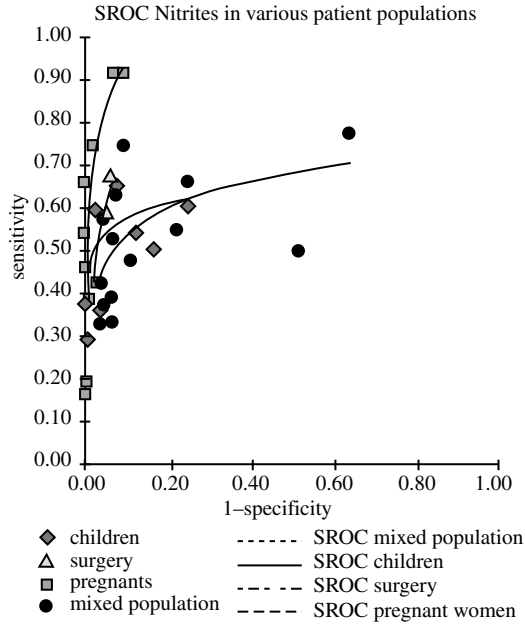


Figure 8.3 Summary ROC curves of nitrites in urine dipsticks for the diagnosis of bacteriuria and urinary tract infections in various homogeneous subgroups of patient populations.

A *random effect model* assumes that in addition to the presence of random error, differences between studies can also result from real differences between study populations and procedures. The weighting factor is mathematically more complex, and is based on the work of Der Simonian and Laird, initially performed and published for the meta-analysis of trials.[38] It includes both within-study and between-study variation.

Homogeneous studies

If the parameters are homogeneous, and if they show no (implicit) cut-off effect, their results can be pooled and a fixed effect model can be used. If there is evidence of a cut-off effect, SROC curves can be constructed or ROC curves can be pooled.

Heterogeneous studies

If heterogeneity is present, the reviewer has the following options:

1. Refrain from pooling and restrict the analysis to a qualitative overview.
2. Subgroup analysis if possible, on prior factors and pooling within homogeneous subgroups.

3. As a last resort pooling can be performed, using methods that are based on a random effect model.

In view of the poor methodological quality of most of the diagnostic studies that have been carried out, there is a tendency to advise using random effect models for the pooling of all diagnostic studies, even if there is no apparent heterogeneity.

Statistical pooling

Pooling of proportions

- *Homogeneous sensitivity and/or specificity*

If *fixed effect* pooling can be used, pooled proportions are the average of all individual study results, weighted for the sample sizes. This is easily done by adding together all numerators and dividing the total by the sum of all denominators[14] (see Appendix to this chapter).

- *Cut-off point effect: SROC curve*

The SROC curve is presented with sensitivity on the y axis and 1 − specificity on the x axis (ROC plot), where each study provides one value of sensitivity and one value of specificity (Figure 8.3). If a SROC curve can be fitted, a regression model (metaregression) is used, with the natural logarithm of the DOR (lnDOR) of the studies as dependent variable and two parameters as independent variables: one for the intercept (to be interpreted as the mean lnDOR) and one for the slope of the curve (as an estimate of the variation of the lnDOR across the studies due to threshold differences). Details and formulae for fitting the curve can be found in the paper presented by Littenberg and Moses[13] (see Appendix). Covariates representing different study characteristics or pretest probabilities can be added to the model to examine any possible association of the diagnostic odds ratio with these variables.[39] The pooled lnDOR and confidence limits have to be back-transformed into a diagnostic odds ratio and its confidence intervals. Metaregression can be unweighted or weighted, using the inverse of the variance as the weighting factor. A problem that is encountered in diagnostic research is the often negative association of the weighting factor with the lnDOR, giving studies with lower discriminative diagnostic odds ratios – because of lower sensitivity and/or specificity – a larger weight. This problem has not yet been resolved.[17,19]

Pooling of likelihood ratios

Continuous test results can be transformed into likelihood ratios, obtained by using different cut-off points. Individual data points from the selected studies can be used to calculate result specific likelihood ratios,[40]

which can be obtained by logistic modelling. The natural log posterior odds are converted into a log likelihood ratio by adding a constant to the regression equation. The constant adjusts for the ratio of the number of "non-diseased" to "diseased" participants in the respective studies[17] (see Appendix).

Pooling of the ROC curves

The results of diagnostic studies with a dichotomous gold standard outcome, and a test result that is reported on a continuous scale, are generally presented as an ROC curve with or without the related area under the curve (AUC) and its 95% CI. To pool such results, the reviewer has three options: to pool sensitivities and specificities for all relevant cut-off points; to pool the AUCs; or to model and pool the ROC curves themselves.

- A pooled ROC curve and its confidence interval can be constructed on the basis of the pooled sensitivity/specificity values per cut-off point. To make this possible, sufficient raw data have to be available, which is seldom the case.
- The AUC, like all one-dimensional measures, provides no information about the asymmetrical nature of a diagnostic test. It cannot distinguish between curves with a high sensitivity at moderate values of the specificity and those with a high specificity at moderate values of the sensitivity.
- As ROC curves are based on ranking, they are robust with respect to interstudy shifts in the value or meaning of cut-off points. They also provide information about the asymmetrical nature of the test information. To enable direct pooling of ROC curves, a method has been developed that requires only the published curves and the number of positive and negative participants on the gold standard test as input.[41] The ROC curve is scanned into a graphic computer file and then converted into a series of sensitivity versus specificity data, using appropiate software or, ultimately, by hand. Subsequently, a model is fitted for each study, similar to that used for producing SROC curves.

  For continuous scale tests, weighted linear regression is used to estimate the parameters for each curve, including a bootstrap method to estimate the standard errors. For ordinal tests, maximum likelihood estimation yields the parameters and their standard errors. The resulting estimates are pooled separately, using a random effect model, and the resulting model is back-transformed into a new pooled curve with its 95% confidence band. In addition to causing calculation problems in specific situations, pooling published ROC curves also hides the test values from the picture. Although this is not a problem when evaluating a test method, or when comparing different methods, it limits the possible use of the pooled curve for evaluating the diagnostic value of

each specific test result. Moreover, a published curve can be a fitted estimate of the real curve based on the initial values, and any bias resulting from this estimation will be included in the pooled estimates.

Data presentation

A DOR is difficult to interpret because it is a combination of sensitivity and specificity. However, it is useful to present pooled sensitivity and specificity estimates, together with the relevant diagnostic odds ratios for different study characteristics or subgroups. To make this information accessible to clinicians, the predictive values could be obtained by using the mean prior (pretest) probabilities of each subgroup. Alternatively, likelihood ratios could be reported so that users can calculate post-test probabilities based on the pretest probabilities applicable to their patients.

Pooled DOR (and confidence intervals) of different subgroups can also be presented graphically on a logarithmic scale to give symmetrical confidence intervals and to reduce the width of confidence intervals.

Example taken from the straight-leg raising test review. In Figure 8.4 the DOR and confidence boundaries are plotted on the *y* axis on a logarithmic scale. Relevant study characteristics (that is, double blind versus single

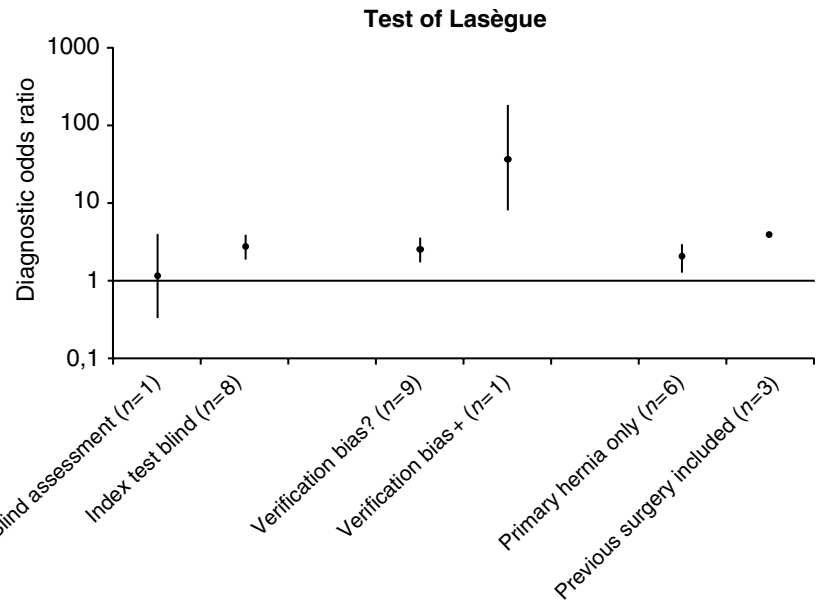


Figure 8.4 Subgroup analyses of the accuracy of Lasègue's test for the diagnosis of disc hernia in low back pain. Odds ratios are pooled per subgroup.

Example taken from the dipstick review:

Factor	DOR (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Prior probability	PPV	NPV
Mixed population	11 (6–21)	0.50 (0.44–0.58)	0.82 (0.71–0.95)	0.32	0.57	0.78
Surgery	34 (25–47)	0.54 (0.39–0.74)	0.96 (0.93–0.99)	0.20	0.76	0.89

blind studies, studies with or without verification bias) are plotted on the *x* axis.

Discussion

Although the methodology to conduct a systematic review and meta-analysis of diagnostic research is developed to a certain extent, at least for dichotomised tests, the exercise itself remains quite a challenge. Systematic reviews have to meet high methodological standards and the results should always be interpreted with caution. Several complicating issues need careful consideration: (1) it is difficult to discover all published evidence, as diagnostic research is often inadequately indexed in electronic databases; (2) the studies are often poorly reported and a set of minimal reporting standards for diagnostic research has only recently been discussed; (3) the methodological quality and validity of diagnostic research reports is often limited (that is, no clear definition of "diseased" participants, no blinding, no independent interpretation of test results, insufficient description of participants); (4) accuracy estimates are often very heterogeneous, yet examining heterogeneity is cumbersome and the process is full of pitfalls; (5) results have to be translated into information that is clinically relevant, taking into account the clinical reality at different levels of health care (prevalence of disease, spectrum of disease, available clinical and other diagnostic information). Even in a state of the art systematic review, the reviewers have to make many subjective decisions when deciding on the inclusion or exclusion of studies, on quality assessment and the interpretation of limited information, on the exclusion of outliers, and on choosing and conducting subgroup analyses. Subjective aspects have to be assessed independently by more than one reviewer, with tracking of disagreements and resolution by consensus or arbitration. These subjective decisions should be explicitly acknowledged in the report to allow the readers some insight into the possible consequences of these decisions on the outcomes of the review and the strength of inference derived from it.

Whereas some researchers question the usefulness of pooling the results of poorly designed research or meta-analyses based on limited information,[42–43] we think that examining the effects of validity criteria on

the diagnostic accuracy measures and the analysis of subgroups adds valuable evidence to the field of diagnostic accuracy studies. The generation of a pooled estimate – the most likely estimate of the test's accuracy – provides clinicians with useful information until better-conducted studies are published. The reader should remember that evidence about the influence of validity of studies on diagnostic accuracy is still limited.[2,3,6] Consequently, it is difficult to recommend a strict set of methodological criteria, recognising that any minimum set of methodological criteria is largely arbitrary. Although we have discussed some practical approaches to statistical pooling, other methods are available in the literature.[18,19] Experience with these methods, however, is limited. The development of guidelines for systematic reviews of tests with continuous or ordinal outcomes, reviews of diagnostic strategies of more than one test, and reviews of the reproducibility of diagnostic tests, remains another challenge, as the methodology is still limited[1] or even non-existent.

References

1 Irwig L, Tosteson ANA, Gatsonis C, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.
2 Lijmer JG, Mol BW, Heisterkamp S, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
3 Oosterhuis WP, Niessen RWLM, Bossuyt PMM. The science of systematic reviewing studies of diagnostic tests. *Clin Chem Lab Med* 2000;**38**:577–88.
4 Jeaschke R, Guyatt GH, Sackett DL. User's guidelines to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA* 1994;**271**:389–91.
5 Jeaschke R, Guyatt GH, Sackett DL. User's guidelines to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA* 1994;**271**:703–7.
6 Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. *BMJ* 1997;**315**:540–3.
7 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* 1995;**274**:645–51.
8 Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in Medline. *J Am Med Informatics Assoc* 1994;**1**:447–58.
9 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994;**309**:1286–91.
10 van der Weijden T, IJzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. *Fam Pract* 1997; **14**:204–8.
11 Devillé WLJM, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000;**53**:65–9.
12 McClish DK. Combining and comparing area estimates across studies or strata. *Med Decision Making* 1992;**12**:274–9.
13 Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decision Making* 1993;**13**:313–21.
14 Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarising diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Med Decision Making* 1993;**13**:253–7.

15 Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293–316.
16 Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psych Bull* 1995; **117**:167–78.
17 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic accuracy. *J Clin Epidemiol* 1995;**48**:119–30.
18 Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995; **2**:S48–S56.
19 Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;**2**:S37–S47.
20 Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;**52**:943–51.
21 Yusuf S, Wittes J, Probsfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. *JAMA* 1991;**266**:93–8.
22 Oxman A, Guyatt G. A consumer's guide to subgroup analysis. *Ann Intern Med* 1992; **116**:78–84.
23 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–5.
24 Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. *Am J Epidemiol* 1995;**142**:371–82.
25 Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. *Ann Intern Med* 1997;**127**:989–95.
26 Devillé WLJM, Yzermans JC, van Duin NP, van der Windt DAWM, Bezemer PD, Bouter LM. Which factors affect the accuracy of the urine dipstick for the detection of bacteriuria or urinary tract infections? A meta-analysis. In: *Evidence in diagnostic research. Reviewing diagnostic accuracy: from search to guidelines*. PhD thesis, Amsterdam: Vrije Universiteit, pp 39–73.
27 Devillé WLJM, van der Windt DAWM, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasègue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;**25**:1140–7.
28 Hunt DL, McKibbon KA. Locating and appraising systematic reviews. *Ann Intern Med* 1997;**126**:532–8.
29 van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM and the Editorial Board of the Cochrane Collaboration Back Review Group. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders. *Spine* 1997;**22**:2323–30.
30 Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Meth Med Res* 1998;**7**:371–92.
31 Grégoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? *J Clin Epidemiol* 1995;**48**:159–63.
32 Cochrane Methods Group on Screening and Diagnostic Tests. Recommended methods. http://www.som.fmc.flinders.edu.au/fusa/cochrane
33 Aertgeerts B, Buntinx F, Kester A, Fevery J. Diagnostic value of the CAGE questionnaire in screening for alcohol abuse and alcohol dependence: a meta-analysis. In: *Screening for alcohol abuse or dependence*. PhD thesis, Leuven: Katholieke Universiteit 2000, Appendix 3.
34 Centre for Evidence Based Medicine. Levels of Evidence and Grades of Recommendations. http://cebm.jr2.ox.ac.uk/docs/levels.html
35 Fleiss Jl. The statistical basis of meta-analysis. *Stat Meth Med Res* 1993;**2**:121–45.
36 Buntinx F, Wauters H. The diagnostic value of macroscopic haematuria in diagnosing urological cancers. A meta-analysis. *Fam Pract* 1997;**14**:63–8.
37 Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;**7**:889–94.
38 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.
39 Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.
40 Irwig L. Modelling result-specific likelihood ratios. *J Clin Epidemiol* 1989;**42**:1021–4.

41 Kester A, Buntinx F. Meta-analysis of ROC-curves. *Med Decision Making* 2000;**20**:430–9.
42 Greenland S. A critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–301.
43 Shapiro S. Meta-analysis/Shmeta-analysis. *Am J Epidemiol* 1994;**140**:771–7.

Appendix: statistical formulae

Pooling of proportions

Homogeneous sensitivity and/or specificity

For example, for the sensitivity:

$$\text{Sensitivity}_{\text{pooled}} = \sum_{i=1}^{k} a_i / \sum (a_i + c_i)$$

where a = true positives, c = false negatives, i = study number, and, k = total number of studies, with standard error:

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}}$$

where $p = \text{sensitivity}_{\text{pooled}}$, and

$$n = \sum_{i=1}^{k} (a_i + c_i)$$

Cut-off point effect: SROC curve

Basic meta-regression formula:

$$\ln(\text{DOR})_{\text{pooled}} = \alpha + \beta S$$

where α = intercept, β = regression coefficient, and

$$S = \text{estimate of cut-off point} = \ln\left[\frac{\text{sensitivity}}{(1 - \text{sensitivity})}\right] + 1\text{n}\left[\frac{(1 - \text{specificity})}{\text{specificity}}\right]$$

With standard error:

$$\text{SE}_{\ln(\text{DOR})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Pooling of likelihood ratios

$$\log(LR)_{\text{pooled}} = \log\left[\frac{N_{\bar{D}}}{N_D}\right] + \alpha + \beta x$$

where LR = likelihood ratio

$$\log\left[\frac{N_{\bar{D}}}{N_D}\right] = \text{correction factor}$$

= log(number non-diseased/number diseased)

α = i ntercept in logistic regression, β = regression coefficient, and x = test measurement.

9 Diagnostic decision support: contributions from medical informatics

JOHAN VAN DER LEI, JAN H VAN BEMMEL

Summary box

- Applying information and communication technology (ICT) to a given medical domain is not merely adding a new technique, but has the potential to radically change processes in that domain.
- When introduced into an environment, ICT will initially often emulate or resemble the existing processes. When workers and researchers in that domain begin to appreciate the potential of ICT, this initial stage is followed by more fundamental changes in that domain that take advantage of the potential of ICT.
- Many researchers argue that the fundamental enabling technology is the introduction of electronic medical records. The explicit purpose of automating medical records is to use the data in those records to support not only the care of individual patients, but also applications such as decision support, quality control, cost control, or epidemiology.
- To understand the scope of the potential changes enabled by electronic records, three principal changes need to be understood. First, data recorded on computer can readily be retrieved and reused for a variety of purposes. Second, once data are available on computer, they can easily be transported. Third, as clinicians (and patients) are using computers to record medical data, the same electronic record can be used to introduce other computer programs that interact with the user.

- As electronic medical records are becoming available, researchers use them to change medical practice by providing decision support, and to analyse observational databases to study the delivery of care. New usage of data, however, generates additional requirements. Thus the experience in developing decision support systems and analysing observational databases feeds back into the requirements for electronic medical records.
- Each patient–doctor encounter, each investigation, each laboratory test, and each treatment in medical practice constitutes, in principle, an experiment. Ideally, we learn from each experiment. Paper as a medium to record data limits our ability to exploit that potential. Electronic medical records will facilitate research that relies on data recorded in routine medical practice. The potential of ICT, however, lies in its ability to close the loop between clinical practice, research, and education.

Introduction

The term medical informatics dates from the second half of the 1970s and is based on the French term *informatique médicale*. Although the term is now widely used, other names, such as medical computer science, medical information sciences, or computers in medicine, are sometimes used. Research in informatics ranges from fundamental computer science to applied informatics. Several definitions of medical informatics take both the scientific, fundamental aspect and the applied, pragmatic aspect into account. Shortliffe, for example, provides the following definition:

"Medical Information Science is the science of using system-analytic tools … to develop procedures (algorithms) for management, process control, decision-making and scientific analysis of medical knowledge."[1] and Van Bemmel defines the field as: "Medical Informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and health care."[1]

Medical informatics is determined by the intersection of the terms medicine and informatics. Medicine identifies the area of research; informatics identifies the methodology used. In medical informatics, we develop and assess methods and systems for the acquisition, processing, and interpretation of patient data. Computers are the vehicles to realise these goals. The role of computers in medical informatics, however, varies. If the medical informatics research is applied, the objective is to develop a computer system that will be used by healthcare professionals, for example

research aimed at the development of electronic medical records. If the research is more fundamental, the computer plays a role as an experimental environment for models that are developed; the objective is not to build a system, but to verify a hypothesis or to investigate the limitations of models. Some research in the area of artificial intelligence in medicine, for example, fits this last category.

Applying ICT to a given medical domain is not merely adding a new technique: ICT has the potential to radically change processes in that domain. Such change, however, may not be apparent at the beginning. When introduced into an environment, ICT will initially often emulate or resemble the already existing processes. Typically, this is only a temporary stage. When workers and researchers in that domain begin to appreciate the potential of ICT, this initial stage is followed by more fundamental changes in that domain that take advantage of the potential of ICT.

Electronic communication, for example, is a relatively simple technology. The contents of a message may be structured (that is, contain a predefined set of data) or free text. When introduced into the healthcare process, electronic communication is used to replace existing paper documents. The names of the first electronic messages often even carry the names of their paper counterpart: electronic discharge letter, electronic prescription, etc. At first glance, little has changed compared to the previous paper based communication, except the speed of delivery. At this stage, the infrastructure (for example computers, lines) required for ICT has been installed, but its impact on the processes is still very limited. Subsequently, however, the ability to send data using electronic communication is used to support new forms of collaboration between healthcare professionals. At present, the emphasis has shifted from replacing paper documents to sharing data between colleagues. As clinicians increasingly share data, issues such as the standardisation of the content of medical records are becoming important areas of research. In addition, the fact that data can be transferred easily over distances enables clinicians to interpret data while the patient is located miles away (resulting in, for example, the so-called "telediagnosis"), or to communicate with patients over longer distances using, for example, the internet.

In this chapter we will focus on the contribution of medical informatics to diagnostic decision support. We believe that the use of ICT in the domain of diagnostic decision support is still in an early stage. Although in a few specialties (for example radiology), ICT is used extensively for decision support, most clinicians have little or no experience with decision support systems. Many researchers argue that the fundamental enabling technology is the introduction of electronic medical records. Once electronic medical records are available, they argue, we will witness a rapid increase in the use of diagnostic decision support systems.[2] We will first discuss the developments with respect to electronic medical records.

Electronic medical records

In its early stages, the written medical record had the purpose of documenting the care given to a patient, and thus to facilitate continuity of that care. The entries in the record enabled the clinician to recall previous episodes of illness and treatment. In recent years, however, medical records have been used increasingly for other purposes: they are used as a data source for purposes ranging from billing the patient to performing epidemiological studies, and from performing quality control to defending oneself against legal claims. One of the major barriers for using the data in such ways is the inaccessible and often unstructured nature of the paper record. The introduction of computer based medical records to a large degree, removes that barrier.

Recent decades have seen a rapid increase in the role of computers in medical record keeping, and professional organisations have started to play an active role in the introduction of electronic records. For example, in 1978 the first Dutch general practitioners started using personal computers in their practices. Five years later, in 1983, 35 general practitioners (that is, 0.6% of all Dutch GPs) were using a computer. In 1990, 35% of Dutch GPs were using one or more computer applications; although the majority of these are administrative, an increasing number of clinicians use computer stored medical records.[3] Now the electronic medical record has replaced paper records as the dominant form of record in Dutch primary care. Other countries, such as the United Kingdom, have also witnessed a rapid introduction of electronic records into primary care. In secondary care, although progress has been made, the introduction of electronic records is slower.

The explicit purpose of automating medical records is to use the data in those records to support not only the care of individual patients, but also applications such as decision support, quality control, cost control, or epidemiology.[2] The quality of medical record data, however, has often been lamentable. The reliability of clinical data, for example, has long been questioned, and tensions between reimbursement schemes and coding schemes have been discussed. Some researchers argue that the process of automation may further reduce the reliability of data. Burnum,[4] for example, states: "With the advent of the information era in medicine, we are pouring out a torrent of medical record misinformation". Although we disagree with this pessimistic view, we acknowledge that medical data are recorded for a specific purpose and that this purpose has an influence on what data are recorded and how. In developing systems that record medical data, designers make decisions about how to model those data in order to perform a given task. For example, in designing the computer based medical record system Elias,[3] the designers focused on issues such as ease of data entry and emulating existing paper records. The same designers

subsequently discovered significant limitations in the Elias records when they developed a decision support system that uses these records as a data source.[5]

Despite the limitations of the current computer based medical records and the data contained in them, many researchers believe electronic medical records will significantly change medical practice.[2,6] To understand the scope of these potential changes, three principles need to be understood. First, data recorded on computer can be readily retrieved and reused for a variety of purposes. As a result, databases containing data on millions of patients are available. Although the subsequent analysis of the data may prove difficult, both clinicians and researchers are moving from a period of "data starvation" to "data overload". Second, once data are available in this way, they can easily be transported. The result is that processes that interpret the data (for example diagnosis or consultation) are no longer closely associated with the physical location where they were collected. Data can be collected in one place and processed in another (for example telediagnosis). Third, as clinicians (and patients) are using computers to record medical data, the same electronic record can be used to introduce other computer programs that interact with the user. Electronic medical records require both an extensive ICT infrastructure and clinicians experienced in using that infrastructure. Once that infrastructure is operational, other applications (such as decision support or access to literature) are much easier to introduce.

Electronic medical records will stimulate and enable other developments. We will discuss two of them: the development and use of integrated decision support systems, and the creation of observational databases.

Clinical decision support systems

A number of definitions for clinical decision support systems have been proposed. Shortliffe,[1] for example, defines a clinical decision support system as: "any computer program designed to help health professionals make clinical decisions".

The disadvantage of such a broad definition is that it includes any program that stores, retrieves or displays medical data or knowledge. To further specify what we mean by the term clinical decision support system, we use the definition proposed by Wyatt and Spiegelhalter[1]: "active knowledge systems that use two or more items of patient data to generate case specific advice". This definition captures the main components of a clinical decision support system: medical knowledge, patient data, and patient specific advice.

In a clinical decision support system, medical knowledge is modelled. That is, the designers of the system encode in a formalism the medical knowledge that is required to make decisions. Such formalisms or models

have traditionally been divided into two main groups: quantitative and qualitative. The quantitative models are based on well defined statistical methods and may rely on training sets of patient data to "train" the model. Examples of such models are neural networks, fuzzy sets, bayesian models, or belief nets. Qualitative models are typically less formal and are often based on perceptions of human reasoning. Examples are truth tables, decision trees, or clinical algorithms. Increasingly, however, builders of decision support systems will combine different models in a given system: a bayesian network may be used to model the diagnostic knowledge and a decision tree may be used to model treatment decisions, and a pharmacokinetic model may be used to calculate dosage regimens.

Clinical decision support systems require patient data. Without them, patient-specific advice cannot be generated. Some systems may require interaction between the system and the clinician. The clinician initiates a dialogue with the system and provides it with data by entering symptoms or answering questions. Experience has shown that the acceptance of this type of system by clinicians is relatively low. Other systems are integrated with electronic medical records, and use the data in them as input. In such settings, receiving decision support requires little or no additional data input on the part of the clinician. Finally, some systems are directly connected to the devices that generate the data, for example systems that interpret ECGs or laboratory data.

By applying the medical knowledge to the patient data, the system generates patient-specific advice. Some systems, especially those integrated with electronic medical records, provide advice independent of a clinician's request for it – unsolicited advice. Examples are reminding systems that continuously screen patient data for conditions that should be brought to the clinician's attention (for example the patient's kidney function is decreasing, or the patient is eligible for preventive screening). Other systems, such as critiquing systems, may monitor the decisions of the clinician and report deviations from guidelines.

Although hundreds of clinical decision support systems have been reported in the literature, only a few have been the subject of a rigorous clinical evaluation. Of those that have been evaluated, however, the majority of the studies showed an impact on clinician performance.[7] In particular, systems integrated with electronic medical records have been demonstrated to improve the quality of care. In light of the currently available evidence, clinical decision support systems constitute a possible method to support the implementation of clinical guidelines in practice.

Observational databases

As a result of the increased use of electronic medical records, large observational databases containing data on millions of patients have

become available for researchers. The data contained in these databases is not of the same quality as the data collected in, for example, a clinical trial. They are collected in routine practice, and, although most observational databases attempt to standardise the recording of data and to monitor their quality, only a few observational databases require the clinician to record additional information.[8]

In the absence of a clear study design (for example a randomised controlled trial to compare the effectiveness of possible treatment regimens) and specification of the data required for that study, the data in observational databases are difficult to interpret because of possible confounding. The advantage of observational databases, however, is that they reflect current clinical practice. Moreover, the data are readily available and the costs are not prohibitive. In settings where all the medical data are recorded electronically, the opportunities for research are similar to studies carried out using paper charts. Compared to paper charts, these observational databases provide an environment where the analysis can be performed faster, the data are legible, "normal" practice can be studied, rare events can be studied, longitudinal follow up of patients is possible, and subgroups for further study (such as additional data collection, or patients eligible for prospective trails) can be identified. In cases where a rapid analysis is required (for example a suspected side effect of a drug), observational databases provide a setting for a quick assessment of the question.

Researchers in medical informatics use these observational databases to assess the behaviour of a decision support system prior to introducing that system into clinical practice. Such an analysis allows the researchers to determine, for example, the frequency of advice (for example the frequency of reminders) and to study the trade off between false positive and false negative alerts. When the clinical decision support system relies on a model that uses patient data to train that model, observational databases allow the tuning of that model to the population in which it will be used.

Observational databases that rely on electronic records have limitations. Analysis of the contents of the records shows that information important for a researcher is often not recorded. Medical records typically contain data describing the patient's state (for example the results of laboratory tests) and the actions of the clinician (such as prescribing medication). Relationships between data are often not recorded. The medical record mainly reflects what is done, rather than why. A further complicating factor is that when data in the medical record describe the relationship between observations (or findings) and actions (for example treatment), the information is often recorded in the form of free text. The clinician's first and most important objective in keeping automated medical records is to document with the purpose of ensuring the quality and continuity of medical care. From the clinician's perspective, free text is often an ideal method for expressing the patient's condition. Researchers, on the other

hand, prefer coded data to facilitate the automatic processing of large numbers of patients. It is unrealistic, however, to expect clinicians to code all relevant data: the time required to do so would render it impractical. In addition, coding is in essence a process of reducing the rich presentation of the patient–doctor encounter to a limited set of predefined terms. The data available in an observational database may therefore not be sufficient to answer a specific research question validly. The completeness of data can only be discussed in the context of a specific study. It is not possible to predict all possible data that would be required for all possible studies. As a result, data in observational databases will be incomplete. Depending on the study question and the impact of incomplete information, additional data may need to be collected.

Confounding by indication

To illustrate the caveats when analysing observational databases, we discuss the problem of confounding by indication. In essence, confounding by indication is often "confounding by diagnosis", since diagnostic assessment is the starting point for any treatment. As an example we will use an observational database used by Dutch GPs: the so-called IPCI database.[8] In this database, we will study the use of long-acting β agonists (LBA). LBA, introduced into the Dutch market in 1992, are long-acting bronchodilators that are used in the treatment of asthma. Dutch guidelines emphasise that LBA are primarily a chronic medication and should be combined with inhaled corticosteroids (IC). We focus on the use of long-acting β_2 agonists and their concomitant use with corticosteroids in general practice.

We conducted a retrospective cohort study during the period 1992–1999 among patients in the Netherlands receiving at least one prescription for one of the long-acting β_2 agonists (LBA) or the short-acting inhaled β_2 agonists. We assessed the indication for the prescription, the characteristics of recipients, the daily dose, and the treatment duration of long-acting β_2 agonists. In addition, we assessed the concomitant use of inhaled corticosteroids, and the incidence of episodes of oral corticosteroid use prior to and during treatment. In the setting of this study, we used the oral corticosteroids as a marker for exacerbations.

We found that the use of LBA among all inhaled β_2 agonist users increased from 3.0% in 1992 to 14.4% in 1999. Of the users, 61% were treated exclusively by the GP and not referred to a specialist. The most common indication for use of LBA was asthma (44%), followed by emphysema (36%) and chronic bronchitis (7%). Only 1.5% of the LBA prescriptions were issued on an "if needed" basis; the most frequent dosing regimen was two puffs per day (78%). Only 67% of the LBA users received inhaled corticosteroids concomitantly. LBA were used for short periods: 32% of the users received only a single prescription, and 49% were treated for less than 90 days.

Table 9.1 For different durations of treatment of patients with asthma with long-acting β_2 agonists (LBA), the frequency of oral corticosteroid use (in prescriptions per 1000 person-days) in the 12 months before the start of the treatment with LBA and during the treatment with LBA.

Duration of treatment with LBA	Corticosteroids 12 months before LBA treatment	Corticosteroids during LBA treatment
From 90 to 180 days	1.3	1.8
From 180 days to 1 year	1.5	1.3
The full year	1.6	0.9

As shown in Table 9.1, among asthma patients who received LBA for at least 1 year the episodes of oral corticosteroid use decreased from 1.6 prescriptions/1000 patient-days (PD) prior to starting LBA, to 0.9/1000 PD during the use of LBA. The rate of corticosteroid use prior to and during LBA use increased in patients with a duration of LBA treatment from 90 to 180 days (from 1.3 to 1.8 per 1000 person-days).

We conclude that the short duration of use of LBAs, and the fact that 33% of patients do not use LBA and IC concomitantly, shows that the use of LBA by Dutch GPs is not in agreement with Dutch guidelines for the treatment of asthma. The interpretation of the use of oral corticosteroids, however, is difficult. The fact that oral corticosteroid use in the patients treated continuously with LBA decreased during treatment, seems to indicate that use of LBA has a positive effect in reducing the number of exacerbations. On the other hand, a comparison of the incidence of exacerbations prior to and during treatment among patients treated with LBA from 90 to 180 days shows that the incidence increases during treatment. It would be dangerous to conclude, based on these crude data, that LBA treatment causes exacerbations in this group of patients. Patients who receive short-term LBA may be different from those with strong fluctuations in asthma severity, for example. The reason (indication) for prescription in this group of patients may be a diagnosed temporary worsening of asthma (increasing severity) that in itself would lead to a higher incidence of exacerbations. Diagnosing severity of the underlying disease that changes over time may therefore, have caused this result: confounded by indication.

In order to be able to adjust for confounding during the analysis of observational studies, we would need an accurate indicator of the severity of the disease over time. For diseases such as asthma it is difficult reliably to assess changing severity by using data collected during routine care. In the absence of a reliable severity indicator any interpretation can be flawed by the potential (residual) confounding. In case of inability to assess severity, the only method to counter confounding would be the randomisation of patients to different treatment arms in order to make both arms similar as to the spectrum of severity. Currently, however, this option is not feasible in “naturalistic” circumstances.

Closing the loop

One of the fundamental changes when conventional paper based medical records are replaced with computerised records involves the ability to process the data in those records for different purposes. As electronic medical records are becoming available, researchers use them to change medical practice by providing decision support, and analyse observational databases to study the delivery of care. New usage of data, however, generates additional requirements. Thus the experience in developing decision support systems and analysing observational databases feeds back into the requirements for electronic medical records. And as new requirements for the electronic record are formulated, the record itself begins to change.

In the area of decision support systems, researchers are combining reminder systems that rely solely on recorded data with systems that request additional information from clinicians. The resulting systems rely on the one hand on data already available in the electronic record to determine eligible patients, and subsequently interact with the clinician to assess, for example, whether the patient should be treated according to a certain protocol. The results of that interaction are recorded in the medical record. Researchers working on the development of observational databases are beginning to combine retrospective research with prospective research. Trials are translated into software, distributed electronically, and added to an electronic medical record. Based on the data in that record, the system automatically detects patients eligible for a trial. It then informs the clinician that the patient is eligible, and requests permission to include them in the trial. The system subsequently performs the randomisation between treatment arms during patient consultation, and the electronic record supports subsequent data collection. As a result, the boundaries between an electronic record, a decision support system, and systems for clinical trials are beginning to fade.

Each patient–doctor encounter, each investigation, each laboratory test, and each treatment in medical practice constitutes, in principle, an experiment. Ideally, we learn from each experiment. Paper as a medium to record data limits our ability to exploit that potential. Electronic medical records will facilitate research that relies on data recorded in routine medical practice. The potential of ICT, however, lies in its ability to close the loop between clinical practice, research, and education.

References

1 van Bemmel JH, Musen MA, eds. *Handbook of Medical Informatics*. Heidelberg: Springer Verlag, 1997.
2 Institute of Medicine, Committee on Improving the Patient Record. *The Computer-Based Patient Record: An Essential Technology for Health Care* (revised edn). Washington DC: National Academy Press, 1997.

3 van der Lei J, Duisterhout JS, Westerhof HP, *et al*. The introduction of computer-based patient records in the Netherlands. *Ann Intern Med* 1993;**119**:1036–41.
4 Burnum JF. The misinformation era: the fall of the medical record. *Ann Intern Med* 1989;**110**:482–4.
5 van der Lei J, Musen MA, van der Does E, Man in't veld AJ, van Bemmel JH. Comparison of computer-aided and human review of general practitioners' management of hypertension. *Lancet* 1991;**338**:1505–8.
6 Knottnerus JA. The role of electronic patient records in the development of general practice in the Netherlands. *Meth Inform Med* 1999;**38**:350–5.
7 Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based decision support systems on clinician performance and patient outcomes: a systematic review. *JAMA* 1998;**280**:1339–46.
8 Vlug AE, van der Lei J, Mosseveld BMT, *et al*. Postmarketing surveillance based on electronic patient records: the IPCI project. *Meth Inform Med* 1999;**38**:339–44.

10 Clinical problem solving and diagnostic decision making: a selective review of the cognitive research literature

ARTHUR S ELSTEIN, ALAN SCHWARTZ

Summary box

- Research on clinical diagnostic reasoning has been conducted chiefly within two research paradigms, problem solving and decision making.
- The key steps in the problem solving paradigm are hypothesis generation, the interpretation of clinical data to test hypotheses, pattern recognition, and categorisation.
- The controversy about whether rapid pattern recognition is accomplished via retrieval of specific instances or by matching to a more abstract prototype can be resolved by recognising that the method selected depends upon the characteristics of the problem.
- Diagnostic performance can be influenced by errors in hypothesis generation and restructuring.
- The decision making paradigm views diagnosis as updating opinion with imperfect information; the normative rule for this process is Bayes' theorem.
- Well documented errors in probability estimation and revision include acquiring redundant evidence, neglecting the role of

disease prevalence in estimating a post-test probability, underestimating the strength of the evidence, the effect of order of presentation of clinical information on final diagnostic conclusions, and the tendency to overestimate the probability of serious but treatable diseases to avoid missing them.
- Problem based learning and ambulatory clinical experiences make sense from the viewpoint of cognitive theory because students generalise less from specific clinical experiences than educators have traditionally hoped.
- A formal quantitative approach to the evidence might have greater generalisability. In residency training, both practice guidelines and evidence-based medicine are seen as responses to the psychological limitations of unaided clinical judgement.

Introduction

This chapter reviews the cognitive processes involved in diagnostic reasoning in clinical medicine and sketches our current understanding of these principles. It describes and analyses the psychological processes and mental structures employed in identifying and solving diagnostic problems of varying degrees of complexity, and reviews common errors and pitfalls in diagnostic reasoning. It does not consider a parallel set of issues in selecting a treatment or developing a management plan. For theoretical background we draw upon two approaches that have been particularly influential in research in this field: problem solving[1–6] and decision making.[7–11]

Problem-solving research has usually focused on how an ill-structured problem situation is defined and structured (as by generating a set of diagnostic hypotheses). Psychological decision research has typically looked at factors affecting diagnosis or treatment choice in well defined, tightly controlled problems. Despite a common theme of limited rationality, the problem-solving paradigm focuses on the wisdom of practice by concentrating on identifying the strategies of experts in a field to help learners acquire them more efficiently. Research in this tradition has aimed at providing students with some guidelines on how to develop their skills in clinical reasoning. Consequently, it has emphasised how experts generally function effectively despite limits on their rational capacities. Behavioural decision research, on the other hand, contrasts human performance with a normative statistical model of reasoning under uncertainty, Bayes' theorem. This research tradition emphasises positive standards for reasoning about uncertainty, demonstrates that even experts in a domain do not always meet

these standards, and thus raises the case for some type of decision support. Behavioural decision research implies that contrasting intuitive diagnostic conclusions with those that would be reached by the formal application of Bayes' theorem would give us greater insight into both clinical reasoning and the probable underlying state of the patient.

Problem solving: diagnosis as hypothesis selection

To solve a clinical diagnostic problem means, first, to recognise a malfunction and then to set about tracing or identifying its causes. The diagnosis is ideally an explanation of disordered function – where possible, a causal explanation. The level of causal explanation changes as fundamental scientific understanding of disease mechanisms evolves. In many instances a diagnosis is a category for which no causal explanation has yet been found.

In most cases, not all of the information needed to identify and explain the situation is available early in the clinical encounter, and so the clinician must decide what information to collect, what aspects of the situation need attention, and what can be safely set aside. Thus, data collection is both sequential and selective. Experienced clinicians execute this task rapidly, almost automatically; novices struggle to develop a plan.

The hypothetico-deductive method

Early hypothesis generation and selective data collection

Difficult diagnostic problems are solved by a process of generating a limited number of hypotheses or problem formulations early in the work up and using them to guide subsequent data collection.[2] Each hypothesis can be used to predict what additional findings ought to be present, if it were true, and then the work up is a guided search for these findings; hence, the method is hypothetico-deductive. The process of problem structuring via hypothesis generation begins with a very limited dataset and occurs rapidly and automatically, even when clinicians are explicitly instructed not to generate hypotheses. Given the complexity of the clinical situation and the limited capacity of working memory, hypothesis generation is a psychological necessity. It structures the problem by generating a small set of possible solutions – a very efficient way to solve diagnostic problems. The content of experienced clinicians' hypotheses are of higher quality; some novices have difficulty in moving beyond data collection to considering possibilities.[3]

Data interpretation

To what extent do the data strengthen or weaken belief in the correctness of a particular diagnostic hypothesis? A bayesian approach to answering

these questions is strongly advocated in much recent writing (for example [12,13]), and is clearly a pillar of the decision making approach to interpreting clinical findings. Yet it is likely that only a minority of clinicians employ it in daily practice, and that informal methods of opinion revision still predominate. In our experience, clinicians trained in methods of evidence-based medicine[14] are more likely to use a bayesian approach to interpreting findings than are other clinicians.

Accuracy of data interpretation and thoroughness of data collection are separate issues. A clinician could collect data thoroughly but nevertheless ignore, misunderstand, or misinterpret some findings. In contrast, a clinician might be overly economical in data collection, but could interpret whatever is available accurately. Elstein et al.[2] found no significant association between thoroughness of data collection and accuracy of data interpretation. This finding led to an increased emphasis upon data interpretation in research and education, and argued for studying clinical judgement while controlling the database. This strategy is currently the most widely used in research on clinical reasoning. Sometimes clinical information is presented sequentially: the case unfolds in a simulation of real time, but the subject is given few or no options in data collection (for example[15–17]). The analysis may focus on memory organisation, knowledge utilisation, data interpretation, or problem representation (for example[3,17,18]). In other studies, clinicians are given all the data simultaneously and asked to make a diagnosis.[19,20]

Pattern recognition or categorisation

Problem-solving expertise varies greatly across cases and is highly dependent on the clinician's mastery of the particular domain. Clinicians differ more in their understanding of problems and their problem representations than in the reasoning strategies employed.[2] From this point of view, it makes more sense to consider reasons for success and failure in a particular case, than generic traits or strategies of expert diagnosticians.

This finding of case specificity challenged the hypothetico-deductive model of clinical reasoning for several reasons: both successful and unsuccessful diagnosticians used hypothesis testing, and so it was argued that diagnostic accuracy did not depend as much on strategy as on mastery of domain content. The clinical reasoning of experts in familiar situations frequently does not display explicit hypothesis testing,[5,21–23] but is instead rapid, automatic, and often non-verbal. The speed, efficiency, and accuracy of experienced clinicians suggests that they might not even use the same reasoning processes as novices, and that experience itself might make hypothesis testing unnecessary.[5] It is likely that experienced clinicians use a hypothetico-deductive strategy only with difficult cases.[24,25] Much of the daily practice of experienced clinicians consists of seeing new cases that strongly resemble those seen previously, and their reasoning in these

situations looks more like pattern recognition or direct automatic retrieval. The question then becomes, what is retrieved? What are the patterns?

Pattern recognition implies that clinical reasoning is rapid, difficult to verbalise, and has a perceptual component. Thinking of diagnosis as fitting a case into a category brings some other issues into clearer view. How is a new case categorised? Two somewhat competing accounts have been offered, and research evidence supports both. Category assignment can be based on matching the case either to a specific instance – so-called instance based or exemplar based recognition – or to a more abstract prototype. In instance based recognition a new case is categorised by its resemblance to memories of instances previously seen.[5,22,26,27] For example, acute myocardial infarction (AMI) is rapidly hypothesised in a 50-year-old male heavy smoker with severe crushing, substernal chest pain because the clinician has seen previous instances of similar men with very similar symptoms who proved to have AMI. This model is supported by the fact that clinical diagnosis is strongly affected by context (for example the location of a skin rash on the body), even when this context is normatively irrelevant.[27] These context effects suggest that clinicians are matching a new case to a previous one, not to an abstraction from several cases, because an abstraction should not include irrelevant features.

The prototype model holds that clinical experience – augmented by teaching, discussion, and the entire round of training – facilitates the construction of abstractions or prototypes.[4,28] Differences between stronger and weaker diagnosticians are explained by variations in the content and complexity of their prototypes. Better diagnosticians have constructed more diversified and abstract sets of semantic relations to represent the links between clinical features or aspects of the problem.[3,29] Support for this view is found in the fact that experts in a domain are more able to relate findings to each other and to potential diagnoses, and to identify the additional findings needed to complete a picture.[24] These capabilities suggest that experts utilise abstract representations and do not merely match a new case to a previous instance.

The controversy about the methods used in diagnostic reasoning can be resolved by recognising that clinicians, like people generally, are flexible in approaching problems: the method selected depends upon the perceived characteristics of the problem. There is an interaction between the clinician's level of skill and the perceived difficulty of the task.[30] Easy cases can be solved by pattern recognition or by going directly from data to diagnostic classification (forward reasoning).[21] Difficult cases need systematic hypothesis generation and testing. Whether a diagnostic problem is easy or difficult is a function of the knowledge and experience of the clinician who is trying to solve it. When we say that a diagnostic problem is difficult, we really mean that a significant fraction of the clinicians who encounter this problem will find it difficult, although for some it may be quite easy.

Errors in hypothesis generation and restructuring

Neither pattern recognition nor hypothesis testing is an error-proof strategy, nor are they always consistent with statistical rules of inference. Errors that can occur in difficult cases in internal medicine are illustrated and discussed by Kassirer and Kopelman.[17] Another classification of diagnostic errors is found in Bordage.[31] The frequency of errors in actual practice is unknown, and studies to better establish the prevalence of various errors are much needed.

Many diagnostic problems are so complex that the correct solution is not contained within the initial set of hypotheses. Restructuring and reformulating occur as data are obtained and the clinical picture evolves. However, as any problem solver works with a particular set of hypotheses, psychological commitment takes place and it becomes more difficult to restructure the problem.[32]

A related problem is that knowledge stored in long-term memory may not be activated unless triggered by a hypothesis or some other cognitive structure that provides an access channel to the contents of memory. This phenomenon has been demonstrated experimentally in a non-clinical context: recall of the details of the layout of a house varies depending on whether one takes the perspective of a burglar or a potential buyer.[33] We are unaware of an experimental demonstration of this effect in medical education, presumably because of the difficulty of ensuring that an experimental trigger has been effective. However, the complaint of many medical educators that students who can solve problems in the classroom setting appear to be unable to do so in the clinic with real patients, illustrates the role of social context in facilitating or hampering access to the memory store. On the other side of this equation there are students who struggle academically but are competent clinicians, presumably because the clinical context facilitates their thinking. These observations are all consistent with Bartlett's[34] classic proposal that memory is organised schematically, not in the storage of unconnected bits. Stories help us to remember the details and also provide guidance as to what details "must be there". This phenomenon has been demonstrated in medical students.[6]

Decision making: diagnosis as opinion revision

Bayes' theorem

From the point of view of decision theory, reaching a diagnosis involves updating an opinion with imperfect information (the clinical evidence).[10–12,35] The normative mathematical rule for this task is Bayes' theorem. The pretest probability is either the known prevalence of the disease or the

clinician's subjective probability of disease before new information is acquired. As new information is obtained, the probability of each diagnostic possibility is continuously revised. The post-test probability – the probability of each disease given the information – is a function of two variables, pretest probability and the strength of the evidence. The latter is measured by a "likelihood ratio", the ratio of the probabilities of observing a particular finding in patients with and without the disease of interest.*

If the data are conditionally independent† each post-test probability becomes the pretest probability for the next stage of the inference process. Using Bayes' theorem becomes hopelessly complicated and impractical when this assumption is violated and more than two correlated cues are involved in a diagnostic judgement, as is often the case in clinical medicine. In these situations, linear and logistic regression techniques are commonly used to derive an equation or clinical prediction rule. A review of these methods is beyond the scope of this chapter. We simply point out that the coefficients (weights) in a regression equation depend on the composition of the derivation sample. Bayes' theorem distinguishes the effect of disease prevalence and the strength of the evidence on the diagnostic judgement, but ordinary regression analytical methods confound these variables in the regression coefficients. (For alternative regression approaches that address this problem, see[36]). If the index disease is overrepresented in the derivation sample, a prediction rule should be applied cautiously to populations where the prevalence of that disease is different. Despite this limitation, these rules are useful. Clinical applications of statistically derived prediction rules can outperform human judgement[37]; this is the rationale for a range of clinical prediction rules that have been developed during the past two decades. Reports of the accuracy of such rules and the reasons for their success have been available in the psychological literature on judgement for over 40 years,[38] but application in clinical practice has been slow because of continuing concerns about:

- whether a rule derived from a particular population generalises accurately to another
- eroding the professional authority and responsibility of clinicians, and
- whether guidelines (at least in the United States) are intended more to ration care and contain costs than to improve quality.[39]

Both evidence-based medicine (EBM) and decision analysis are efforts to introduce quantification into the diagnostic process and still leave a

* Formally, LR+ = Sensitivity/(1−Specificity) and LR− = (1−Sensitivity)/Specificity.

† Two tests are conditionally independent if the sensitivity and specificity of each test is the same whether the other is positive or negative. Formally, if T_1 and T_2 are conditionally independent tests for a disease D, then:
$p(T_2+ \mid D+ \text{ and } T_1+) = p(T_2+ \mid D+)$ and
$p(T_2- \mid D- \text{ and } T_{1-}) = p(T_2- \mid D-)$

substantial role for clinical judgement.[40,41] EBM leaves the application of research results, including a clinical guideline, up to the clinical judgement of the clinician, who should be guided by canons for interpreting the literature. Decision analysis proposes to offer the clinician insight into the crucial variables in a decision problem, together with a recommended strategy that maximises expected utility (for example, see[42]). Both attempt to avoid quasimandatory prescriptive guidelines and to leave room for professional discretion.

Bayes' theorem is a normative rule for diagnostic reasoning: it tells us how we *should* reason, but it does not claim that we use it to revise opinion. It directs attention to two major classes of error in clinical reasoning: in the assessment of either pretest probability or the strength of the evidence. The psychological study of diagnostic reasoning from the bayesian viewpoint has focused on errors in both components.

Errors in probability estimation

Availability

People are prone to overestimate the frequency of vivid or easily recalled events and to underestimate the frequency of events that are either very ordinary or difficult to recall.[43,44] Diseases or injuries that receive considerable media attention (for example injuries due to shark attacks) are often considered more probable than their true prevalence. This psychological principle is exemplified clinically in *overemphasising rare conditions*. Unusual cases are more memorable than routine problems. The clinical aphorism "When you hear hoofbeats, think horses, not zebras" calls attention to this bias.

Representativeness

Earlier, clinical diagnosis was viewed as a categorisation process. The strategy of estimating the probability of disease by judging how similar a case is to a diagnostic category or prototype can lead to an overestimation of the probability of a disease in two ways. First, post-test probability can be confused with test sensitivity.[45,46] For example, although fever is a typical finding in meningitis, the probability of meningitis given fever alone as a symptom is quite low. Second, representativeness neglects base rates and implicitly considers all hypotheses as equally likely. This is an error, because if a case resembles disease A and disease B equally well, and there are 10 times as many cases of A as of B, then the case is more likely an instance of A. This heuristic drives the "conjunction fallacy": incorrectly concluding that the probability of a joint event (such as the combination of multiple symptoms to form a typical clinical picture) is greater than the probability of any one of those events alone. The joint event may be more representative (typical) of the diagnostic category, but it cannot be more probable than a single component.

Probability distortions

Normative decision theory assumes that probabilities are mentally processed linearly, that is, they are not transformed from the ordinary probability scale. Because behavioural decision research has demonstrated several violations of this principle, it has been necessary to formulate descriptive theories of risky choice that will better account for choice behaviour in a wide range of situations involving uncertainty. One of the earliest of these theories is prospect theory (PT),[47] which was formulated explicitly to account for choices involving two-outcome gambles (or one two-outcome gamble and a certain outcome). Cumulative prospect theory (CPT)[48] extends the theory to the multioutcome case. Both PT and CPT propose that decision makers first edit the decision stimulus in some way, and then evaluate the edited stimulus. Options are evaluated by using an expected-utility-like rule, except that a transformation of the probabilities, called decision weights, are multiplied by subjective values and summed to yield the valuation of a lottery. Probabilities are transformed by a function that is sensitive to both the magnitude of each probability and its rank in the cumulative probability distribution. Hence, it is a *rank-dependent* utility theory. In general, small probabilities are overweighted and large probabilities underweighted. This "compression error"[49] results in discontinuities at probabilities of 0 and 1, and permits this model to predict "certainty effect" violations of expected utility theory (in which the difference between 99% and 100% is psychologically much greater than the difference between, say, 60% and 61%). Cumulative prospect theory and similar rank-dependent utility theories provide formal descriptions of how probabilities are distorted in risky decision making. The distortions are exacerbated when the probabilities are not precisely known,[50] a situation that is fairly common in clinical medicine. It should be stressed that cumulative prospect theory does not assert that individuals are in fact carrying out mentally a set of calculations which are even more complex than those required to calculate expected utility. Rather, the theory claims that observed choices (that is, behaviour) can be better modelled by this complex function than by the simpler expected-utility rule.

Support theory

Several probability estimation biases are captured by support theory,[51–53] which posits that subjective estimates of the frequency or probability of an event are influenced by how detailed the description is. More explicit descriptions yield higher probability estimates than compact, condensed descriptions, even when the two refer to exactly the same events (such as "probability of death due to a car accident, train accident, plane accident, or other moving vehicle accident" versus "probability of death due to a moving vehicle accident"). This theory can explain availability (when

memories of an available event include more detailed descriptions than those of less available events) and representativeness (when a typical case description includes a cluster of details that "fit", whereas a less typical case lacks some of these features). Clinically, support theory implies that a longer, more detailed case description will be assigned a higher subjective probability of the index disease than a brief abstract of the same case, even if they contain the same information about that disease. Thus, subjective assessments of events, although often necessary in clinical practice, can be affected by factors unrelated to true prevalence.[53]

Errors in probability revision

Errors in interpreting the diagnostic value of clinical information have been found by several research teams.[2,6,54,55]

Conservatism

In clinical case discussions data are commonly presented sequentially, and diagnostic probabilities are not revised as much as is implied by Bayes' theorem. This "stickiness" has been called "conservatism", and was one of the earliest cognitive biases identified.[56]

Anchoring and adjustment

One explanation of conservatism is that diagnostic opinions are revised up or down from an initial anchor, which is either given in the problem or formed subjectively. Anchoring and adjustment means that final opinions are sensitive to the starting point (the "anchor"), that the shift ("adjustment") from it is typically insufficient, and so the final judgement is closer to the anchor than is implied by Bayes' theorem.[43] Both of these biases will lead to the collection of more information than is normatively necessary to reach a desired level of diagnostic certainty. The common complaint that clinicians overuse laboratory tests is indirect evidence that these biases operate in clinical practice.

Confounding the probability and value of an outcome

It is difficult for everyday judgement to keep separate accounts of the probability of a particular disease and the benefits that accrue from detecting it. Probability revision errors that are systematically linked to the perceived cost of mistakes demonstrate the difficulties experienced in separating assessments of probability from values.[57,58] For example, there is a tendency to overestimate the probability of more serious but treatable diseases, because a clinician would hate to miss one.[57]

Acquiring redundant evidence

"Pseudodiagnosticity"[59] or "confirmation bias"[54] is the tendency to seek information that confirms a hypothesis rather than the data that facilitate efficient testing of competing hypotheses. For example, in one study, residents in internal medicine preferred about 25% of the time to order findings that would give a more detailed clinical picture of one disease, rather than findings that would allow them to test between two potential diagnoses.[54] Here, the problem is knowing what information would be useful, rather than overestimating the value (likelihood ratio) of the information, or failing to combine it optimally with other data.[55]

Incorrect interpretation

A common error is interpreting findings as consistent with hypotheses already under consideration.[2,6,60] Where findings are distorted in recall, it is generally in the direction of making the facts more consistent with typical clinical pictures.[2] Positive findings are overemphasised and negative findings tend to be discounted.[2,61] From a bayesian standpoint these are all errors in assessing the diagnostic value of information, that is, errors in subjectively assessing the likelihood ratio. Even when clinicians agree on the presence of certain clinical findings, wide variations have been found in the weights assigned in interpreting cues,[20] and this variation may be due partly to the effect of the hypotheses being considered.[60]

Order effects

Bayes' theorem implies that clinicians given identical information should reach the same diagnostic opinion, regardless of the order in which the information is presented. However, final opinions are also affected by the order of presentation: information presented later in a case is given more weight than that presented earlier.[15,62] This may partly explain why it is difficult to get medical students to pay as much attention to history and the physical examination as their teachers would wish. Modern diagnostic studies tend to have very high likelihood ratios, and they also are obtained late in the diagnostic work up.

Educational implications

Two recent innovations in undergraduate medical education and residency training, problem based learning and evidence-based medicine, are consistent with the educational implications of this research.

Problem based learning (PBL)[63–65] can be understood as an effort to introduce formulating and testing clinical hypotheses into a preclinical curriculum dominated by biological sciences. The cognitive–instructional

theory behind this reform was that, because experienced clinicians use this strategy with difficult problems, and as practically any clinical situation selected for instructional purposes will be difficult for students, it makes sense to call this strategy to their attention and to provide opportunities to practise it, first using case simulations and then with real patients.

The finding of case specificity showed the limits of a focus on teaching a general problem solving strategy. Problem solving expertise can be separated from content analytically, but not in practice. This realisation shifted the emphasis toward helping students acquire a functional organisation of content with clinically usable schemata. This became the new rationale for problem based learning.[66,67]

Because transfer from one context to another is limited, clinical experience is needed in contexts closely related to future practice. The instance-based model of problem solving supports providing more experience in outpatient care because it implies that students do not generalise as much from one training setting to another as has traditionally been thought. But a clinician overly dependent on context sees every case as unique, as all of the circumstances are never exactly replicated. The unwanted intrusion of irrelevant context effects implies that an important educational goal is to reduce inappropriate dependence on context. In our opinion, there are two ways to do this:

1. Emphasise that students should strive to develop prototypes and abstractions from their clinical experience. Clinical experience that is not subject to reflection and review is not enough. It must be reviewed and analysed so that the correct general models and principles are abstracted. Most students do this, but some struggle, and medical educators ought not to count upon its spontaneous occurrence. Well designed educational experiences to facilitate the development of the desired cognitive structures should include extensive focused practice and feedback with a variety of problems.[5,68] The current climate, with its emphasis on seeing more patients to compensate for declining patient care revenues, threatens medical education at this level because it makes it more difficult for clinical preceptors to provide the needed critique and feedback, and for students to have time for the necessary reflection.[69]
2. A practical bayesian approach to diagnosis can be introduced. EBM[14,70] may be viewed as the most recent – and, by most standards, the most successful – effort to date to apply these methods to clinical diagnosis. EBM uses likelihood ratios to quantify the strength of the clinical evidence, and shows how this measure should be combined with disease prevalence (or prior probability) to yield a post-test probability. This is Bayes' theorem offered to clinicians in a practical, useful format! Its strengths are in stressing the role of data in clinical reasoning, and in

encouraging clinicians to rely on their judgement to apply the results of a particular study to their patients. Its weaknesses, in our view, are that it does not deal systematically with the role of patient preferences in these decisions, or with methods for quantifying preferences, and that it blurs the distinction between probability driven and utility driven decisions.

In our experience teaching EBM, residents soon learn how to interpret studies of diagnostic tests and how to use a nomogram[70,71] to compute post-test probabilities. The nomogram, or a 2×2 table, combines their prior index of suspicion (a subjective probability) and the test characteristics reported in the clinical literature. It has been more difficult to introduce concepts of decision thresholds (at what probability should management change?) and the expected value of information (should a test that cannot result in a change in action be performed at all?).

Methodological guidelines

1. Psychological research on clinical reasoning began in a thinking-aloud tradition, which remains attractive to many investigators. It seems quite natural to ask a clinician to articulate and discuss the reasoning involved in a particular case, and to record these verbalisations for later analysis. Whatever its shortcomings, this research strategy has high face validity. Because the clinicians involved in these studies frequently discuss real cases (for example[2,17]), content validity on the clinical side is not a problem.

The problems of this approach are easily summarised: first, it is labour intensive, and therefore most studies have used small samples of both clinicians and cases. Therefore, they lack statistical power and are best suited for exploratory analysis. But to demonstrate statistically significant differences between experts (senior attending clinicians) and novices (medical students or junior house officers), researchers must take into account two facts: (1) within any group of clinicians at any level of clinical experience, or within any speciality, there is a great amount of variation, both in reasoning and in practice. With small samples, within-group variance will make it difficult to demonstrate significant between-group differences; and (2) the performance of clinicians varies considerably across cases. These two features imply that research on diagnostic reasoning must use adequate samples of both clinicians and cases if there is to be any hope of reaching generalisable conclusions. Most research to date has not paid adequate attention to issues of sample size (of both cases and research participants) and statistical power.

2. Many important cognitive processes are not available to consciousness and are not verbalised. Indeed, the more automatic and overlearned a mental process is, the less likely is it that one can verbalise

how the process works. Once a person has lived for some time at a given address, it becomes impossible to tell how one knows that address: it is simply "known". Closer to our concerns, participants do not report that a subjective probability is overestimated because of the availability bias: the existence of the bias is inferred by comparing estimates with known frequencies. For these reasons, much recent work has shifted toward a research paradigm that owes much to experimental cognitive psychology: research participants are presented with a task and their responses are recorded. Their verbalisations are one more source of data, but are not treated as a true account of internal mental processes. This research has yielded many of the findings summarised in this chapter, but it is at times criticised for using artificial tasks (lack of face validity) and, consequently, not motivating the participants adequately. The generalisability of the results to real clinical settings is then questioned.

3. Selection bias is a potential threat to the validity of both types of studies of clinical reasoning. Senior clinicians in any clinical domain can decline to participate in research far more easily than can medical students or house officers in the same domain. Therefore, the more experienced participants in a study are usually volunteers. Attention should be paid to issues of selection bias and response rate as potential limitations; thought should be given to their possible effects on the validity and generalisability of the results of the study.

4. Behavioural decision research conducted to date has been concerned primarily with demonstrating that a particular phenomenon exists, for example demonstrating biases in probability estimation, such as availability and representativeness. Statistical tests of significance are used to demonstrate the phenomena. From an educational standpoint, we ought to be more interested in identifying how prevalent these biases are and which are most likely to affect treatment and management. Thus, more research is needed to assess the prevalence of these errors and to determine how often treatment choices are affected by diagnostic errors caused by these biases. If these facts were known, a more rational, systematic curriculum could be developed.

Conclusion

This chapter has selectively reviewed 30 years of psychological research on clinical diagnostic reasoning, focusing on problem solving and decision making as the dominant paradigms of the field. This research demonstrates the limitations of human judgement, although the research designs employed make it difficult to estimate their prevalence. Problem based learning and evidence-based medicine are both justified by the psychological research about judgement limitations, violations of bayesian principles in everyday clinical reasoning, and the finding of limited transfer

across clinical situations, although we do not believe that these innovations were initially directed by an awareness of cognitive limitations. Within graduate medical education (residency training), the introduction of practice guidelines based on evidence has been controversial because guidelines may be perceived as efforts to restrict the authority of clinicians and to ration care. The psychological research helps to explain why formal statistical decision supports are both needed and likely to evoke controversy.

Preparation of this review was supported in part by grant RO1 LM5630 from the National Library of Medicine.

References

1 Newell A, Simon HA. *Human problem solving*. Englewood Cliffs (NJ): Prentice-Hall, 1972.
2 Elstein AS, Shulman LS, Sprafka SA. *Medical problem solving: An analysis of clinical reasoning*. Cambridge MA: Harvard University Press, 1978.
3 Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. *Acad Med* 1991;**66**(9 Suppl):S70–S72.
4 Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. *Med Educ* 1984;**18**:406–16.
5 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med* 1990;**65**:611–21.
6 Friedman MH, Connell KJ, Olthoff AJ, Sinacore J, Bordage G. Medical student errors in making a diagnosis. *Acad Med* 1998;**73**(10 Suppl):S19–S21.
7 Kahneman D, Slovic P, Tversky A (eds.) *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press, 1982.
8 Baron J. *Thinking and deciding*. New York: Cambridge University Press, 1988.
9 Mellers BA, Schwartz A, Cooke ADJ. Judgment and decision making. *Annu Rev Psychol* 1998;**49**:447–77.
10 Weinstein MC, Fineberg HV, Elstein AS, *et al*. *Clinical decision analysis*. Philadelphia: WB Saunders, 1980.
11 Sox HC Jr, Blatt MA, Higgins MC, Marton KI. *Medical decision making*. Stoneham (MA): Butterworths, 1988.
12 Pauker SG. Clinical decision making: handling and analyzing clinical data. In: Bennett JC, Plum F, eds. *Cecil's Textbook of Medicine*, 20th edn. Philadelphia: WB Saunders, 1996;78–83.
13 Panzer RJ, Black ER, Griner PF. *Diagnostic strategies for common medical problems*. Philadelphia: American College of Physicians, 1991.
14 Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. New York: Churchill Livingstone, 1997.
15 Chapman GB, Bergus GR, Elstein AS. Order of information affects clinical judgment. *J Behav Decision Making* 1996;**9**:201–11.
16 Moskowitz AJ, Kuipers BJ, Kassirer JP. Dealing with uncertainty, risks, and tradeoffs in clinical decisions: a cognitive science approach. *Ann Intern Med* 1988;**108**:435–49.
17 Kassirer JP, Kopelman RI. *Learning clinical reasoning*. Baltimore: Williams & Wilkins, 1991.
18 Joseph GM, Patel VL. Domain knowledge and hypothesis generation in diagnostic reasoning. *Med Decision Making* 1990;**10**:31–46.
19 Patel VL, Groen G. Knowledge-based solution strategies in medical reasoning. *Cogn Sci* 1986;**10**:91–116.
20 Wigton RS, Hoellerich VL, Patil KD. How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. *Med Decision Making* 1986;**6**:2–11.

21 Groen GJ, Patel VL. Medical problem-solving: some questionable assumptions. *Med Educ* 1985;**19**:95–100.
22 Brooks LR, Norman GR, Allen SW. Role of specific similarity in a medical diagnostic task. *J Exp Psych: Gen* 1991;**120**:278–87.
23 Eva KW, Neville AJ, Norman GR. Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Acad Med* 1998;**73**(10 Suppl):S1–S5.
24 Elstein AS, Kleinmuntz B, Rabinowitz M, *et al*. Diagnostic reasoning of high- and low-domain knowledge clinicians: a re-analysis. *Med Decision Making* 1993;**13**:21–9.
25 Davidoff F. *Who has seen a blood sugar? Reflections on medical education*. Philadelphia: American College of Physicians, 1998.
26 Medin DL, Schaffer MM. A context theory of classification learning. *Psychol Rev* 1978;**85**:207–38.
27 Norman GR, Coblentz CL, Brooks LR, Babcock CJ. Expertise in visual diagnosis: a review of the literature. *Acad Med* 1992;**66**:S78–S83.
28 Rosch E, Mervis CB. Family resemblances: studies in the internal structure of categories. *Cogn Psychol* 1975;**7**:573–605.
29 Lemieux M, Bordage G. Propositional versus structural semantic analyses of medical diagnostic thinking. *Cogn Sci* 1992;**16**:185–204.
30 Elstein AS. What goes around comes around: the return of the hypothetico-deductive strategy. *Teach Learn Med* 1994;**6**:121–3.
31 Bordage G. Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Acad Med* 1999;**74**:S138–S142.
32 Janis IL, Mann L. *Decision-making*. New York: Free Press, 1977.
33 Anderson RC, Pichert JW. Recall of previously unrecallable information following a shift in perspective. *J Verb Learning Verb Behav* 1978;**17**:1–12.
34 Bartlett FC. *Remembering: a study in experimental and social psychology*. New York: Macmillan, 1932.
35 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: A basic science for clinical medicine*, 2nd edn. Boston: Little, Brown and Company; 1991.
36 Knottnerus JA. Application of logistic regression to the analysis of diagnostic data. *Med Decision Making* 1992;**12**:93–108.
37 Ebell MH. Using decision rules in primary care practice. *Prim Care* 1995;**22**:319–40.
38 Meehl PE. *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press, 1954.
39 James PA, Cowan TM, Graham RP, Majeroni BA. Family physicians' attitudes about and use of clinical practice guidelines. *J Fam Pract* 1997;**45**:341–7.
40 Hayward R, Wilson MC, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA* 1995;**274**:70–4.
41 Wilson MC, Hayward R, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help me in caring for my patient? *JAMA* 1995;**274**:1630–2.
42 Col NF, Eckman MH, Karas RH, *et al*. Patient-specific decisions about hormone replacement therapy in postmenopausal women. *JAMA* 1997;**277**:1140–7.
43 Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;**185**:1124–31.
44 Elstein AS. Heuristics and biases: selected errors in clinical reasoning. *Acad Med* 1999;**74**:791–4.
45 Eddy DM. Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press; 1982;249–67.
46 Dawes, RM. *Rational choice in an uncertain world*. New York: Harcourt Brace Jovanovich, 1988.
47 Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1982;**211**:453–8.
48 Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;**5**:297–323.

49 Fischhoff B, Bostrom A, Quadrell MJ. Risk perception and communication. *Annu Rev Pub Health* 1993;**14**:183–203.
50 Einhorn HJ, Hogarth RM. Decision making under ambiguity. *J Business* 1986; **59**(Suppl):S225–S250.
51 Tversky A, Koehler DJ. Support theory: a nonextensional representation of subjective probability. *Psychol Rev* 1994;**101**:547–67.
52 Rottenstreich Y, Tversky A. Unpacking, repacking, and anchoring: advances in support theory. *Psychol Rev* 1997;**104**:406–15.
53 Redelmeier DA, Koehler DJ, Liberman V, Tversky A. Probability judgment in medicine: discounting unspecified probabilities. *Med Decision Making* 1995;**15**:227–30.
54 Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing hypotheses heuristic: A practical approach to judgment under uncertainty and Bayesian probability. *JAMA* 1985;**253**:2858–62.
55 Gruppen LD, Wolf FM, Billi JE. Information gathering and integration as sources of error in diagnostic decision making. *Med Decision Making* 1991;**11**:233–9.
56 Edwards W. Conservatism in human information processing. In: Kleinmuntz B, ed. *Formal representation of human judgment*. New York, Wiley, 1969. pp. 17–52.
57 Wallsten TS. Physician and medical student bias in evaluating information. *Med Decision Making* 1981;**1**:145–64.
58 Poses RM, Cebul RD, Collins M, Fager SS. The accuracy of experienced physicians' probability estimates for patients with sore throats. *JAMA* 1985;**254**:925–9.
59 Kern L, Doherty ME. "Pseudodiagnosticity" in an idealized medical problem-solving environment. *J Med Educ* 1982;**57**:100–4.
60 Hatala R, Norman GR, Brooks LR. Influence of a single example on subsequent electrocardiogram interpretation. *Teach Learn Med* 1999;**11**:110–17.
61 Wason PC, Johnson-Laird PN. *Psychology of reasoning: Structure and content*. Cambridge (MA): Harvard University Press, 1972.
62 Bergus GR, Chapman GB, Gjerde C, Elstein AS. Clinical reasoning about new symptoms in the face of pre-existing disease: sources of error and order effects. *Fam Med* 1995; **27**:314–20.
63 Barrows HS. Problem-based, self-directed learning. *JAMA* 1983;**250**:3077–80.
64 Barrows HS. A taxonomy of problem-based learning methods. *Med Educ* 1986;**20**:481–6.
65 Schmidt HG. Problem-based learning: rationale and description. *Med Educ* 1983; **17**:11–16.
66 Norman GR. Problem-solving skills, solving problems and problem-based learning. *Med Educ* 1988;**22**:279–86.
67 Gruppen LD. Implications of cognitive research for ambulatory care education. *Acad Med* 1997;**72**:117–20.
68 Bordage G. The curriculum: overloaded and too general? *Med Educ* 1987;**21**:183–8.
69 Ludmerer KM. *Time to heal: American medical education from the turn of the century to the era of managed care*. New York: Oxford University Press, 1999.
70 Fagan TJ. Nomogram for Bayes' theorem. *N Engl J Med* [Letter] 1975;**293**:257.
71 Schwartz A. Nomogram for Bayes' theorem. Available at http://araw.mede.uic.edu/cgi-bin/testcalc.pl.

11 Improving test ordering and diagnostic cost effectiveness in clinical practice – bridging the gap between clinical research and routine health care

RON AG WINKENS, GEERT-JAN DINANT

Summary box

- In recent decades the number of diagnostic tests ordered by doctors has increased enormously, despite the often absent or disappointing results from studies into their accuracy.
- Contradictions between scientific evidence and daily practice can be obstacles to improving test ordering behaviour.
- Evidence-based clinical guidelines are needed to formalise optimal diagnostic performance, but will not work unless implemented properly.
- There is no "one and only ideal implementation strategy". Often, a combination of supportive interventions is the best approach.
- Interventions should provide both knowledge on what to do and insight into one's own performance. Such interventions are audit, individual feedback, peer review, and computer reminders.
- From a viewpoint of cost effectiveness, computer interventions look promising.

- More attention must be paid to the perpetuation of interventions once they have been started, and to the measurement and scientific evaluation of their effects over time.
- Although the randomised controlled trial remains the "gold standard" for evaluation studies, inherent methodological challenges, such as the required randomisation of doctors, need special attention.
- More research into the ways to improve test ordering is urgently needed, in particular for patients suffering from recurrent non-specific or unexplained complaints.

Introduction

An important part of making a proper diagnosis is using diagnostic tests, such as blood and radiographic investigations. In recent decades the total number of diagnostic tests ordered by doctors has increased substantially, despite the often disappointing results from studies into their diagnostic accuracy. Apparently, arguments other than scientific ones for test ordering are relevant. Furthermore, it might be questioned to what extent current knowledge, insights into the correct use of diagnostic tests, and results from research have been properly and adequately implemented in daily practice. This chapter will discuss how to bridge the gap between evidence from clinical research and routine health care.

The need to change

For several reasons there is a need to improve test ordering behaviour. The use of medical resources in western countries is growing annually and consistently. In the Netherlands, for example, there is a relatively stable growth in nationwide expenditures for health care of approximately 7% per year. The growth in expenditure for diagnostic tests is similar. However, whereas expenditure increases, health status does not seem to improve accordingly. This at least suggests that there is a widespread and general overuse of diagnostic tests.

The following factors may be responsible for the increasing use of diagnostic tests. First, the mere availability and technological imperative of more test facilities is an important determinant. In view of the interaction between supply and demand in health care, the simple fact that tests can be ordered will lead to their actual ordering. This applies especially to new tests, which are sometimes used primarily out of curiosity. Another factor is the increasing demand for care, caused partly by the ageing of the

population and an increasing number of chronically ill people. Also, new insights from scientific evidence and guidelines often provide recommendations for additional diagnostic testing.

Doctors might wish to perform additional testing once an abnormal test result is found, even in the absence of clinical suspicion, while ignoring that a test result outside reference ranges may generally be found in 50% of a healthy population. A cascade of testing may then be the result.

Furthermore, over the years, higher standards of care (adopted by the public, patients, and healthcare professionals) and defensive behaviours from doctors have contributed to the increased use of healthcare services, one of them being diagnostic testing.

In summary, despite the introduction of guidelines focusing on rational use of diagnostic tests, in daily practice reasons for ignoring evidence-based recommendations (such as fear of patients, and the doctor's wish to gain time) are numerous and hard to grasp.

Altogether, the ensuing problem is threefold. First, there are reasons to believe that some of the tests requested are non-rational. Second, contradictions between scientific evidence and daily practice can be ultimate obstacles to changing test ordering behaviour. And third, there is an increasing tension between volume growth and financial constraints. With this in mind, it is clear that there is a need to achieve a more efficient use of diagnostic tests. If interventions meant to put a stop to the unbridled growth in the number of tests ordered were to focus especially on situations where their use is clearly inappropriate, healthcare expenditure might be reduced and the quality of care might even improve.

Can we make the change?

In terms of quality improvement and cost containment there are sufficient arguments for attempting to change test ordering behaviour. To do so, it is recommended that certain steps be taken, from orientation to perpetuation. The individual steps are described in the implementation cycle[1] (Figure 11.1).

Following the implementation cycle, several things need to be done. First, insight into the problem is needed. The problem needs to be well defined and must be made clear to those whose performance we wish to change. Next, the optimal – "gold standard" – situation should be determined, and communicated as such. Usually this means the development and dissemination of guidelines. Also, an assessment of actual performance (the level of actual care) is needed. Then, the desired changes need to be determined and an implementation strategy set up to achieve the actual change. After this, the results should be monitored. The outcome of this monitoring can be used as new input for further improvement and for defining new goals for quality assurance, thereby re-entering the implementation cycle. It should be noted that these general rules apply

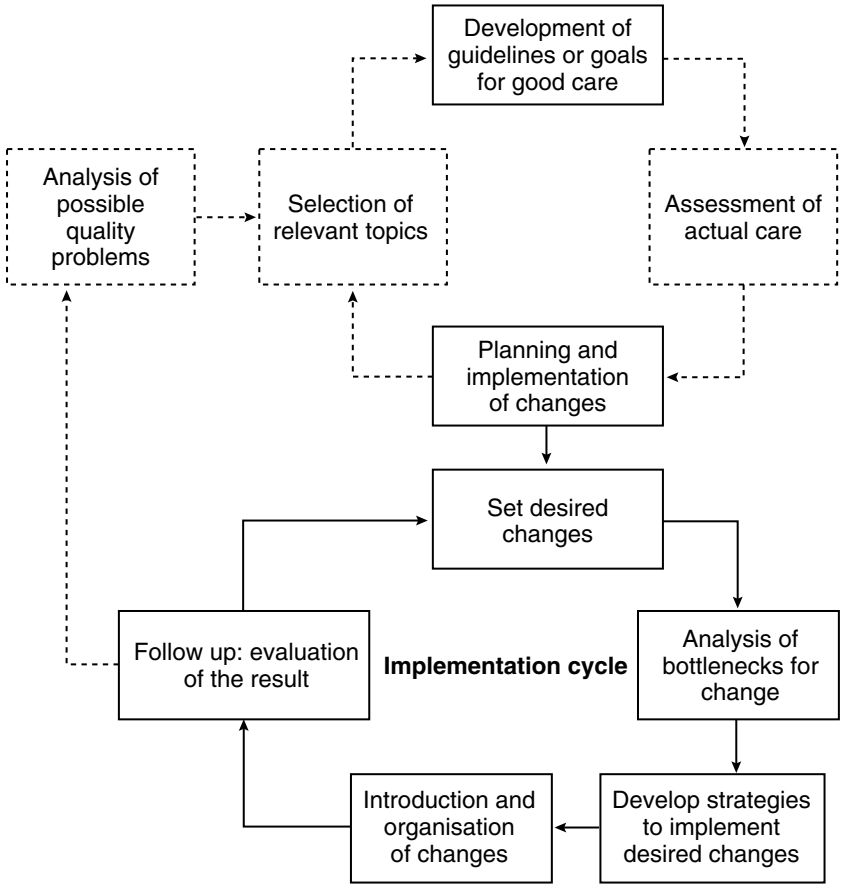


Figure 11.1 The implementation cycle.[1]

not only to test ordering behaviour, but also to other actions (such as prescribing drugs and referring to hospital care). The following highlights a number of steps in the implementation cycle.

If we are to change the test ordering behaviour of clinicians the first move is to assess and establish what the desired optimal situation would be and how the actual performance should look. Guidelines, protocols and standards are needed to formalise this optimal situation. In the past there have been various moves toward guideline development. By and large, these guidelines are problem oriented and address clinical problems as a whole, such as taking care of diabetes patients, or diagnosing and managing dyspepsia. The diagnostic work up of a clinical problem and the resulting recommendations for specific tests – if necessary – are then important aspects to be dealt with. A good example of a comprehensive set of

guidelines is the development of standards for Dutch GPs by the Dutch College of General Practitioners.[2] Starting in 1989, the College has set up almost 70 guidelines on a variety of common clinical problems. In the meantime, many of these standards have been revised in line with new evidence from scientific research. One of the guidelines specifically addresses the principles of rational ordering of blood tests.[3]

The development of guidelines does not automatically lead to the desired behavioural change, especially when their dissemination is limited to the distribution of written material. In other words, simply publishing and mailing the guidelines does not make clinicians act accordingly. Implementation strategies are needed to bring about actual change. Implementation actually includes a whole range of activities to stimulate the use of guidelines. Such activities may include communication and information, giving insight into the problem and the need to change, and specific activities to achieve the actual change in behaviour.

One of the oldest and most frequently used interventions is postgraduate continuous medical education (CME), varying from lectures to comprehensive courses and training sessions. Nowadays, CME is increasingly connected to (re)certification. Another intervention that can be considered an implementation strategy is written material through books, papers in the literature, and protocol manuals. Although part of the written material is intended as a way to distribute research findings and to increase the level of up to date scientific knowledge, among clinicians it is often also focused as improving clinical performance. There is a range of interventions that combine the provision of knowledge with giving more specific insight into one's own performance. Such interventions include audit, individual feedback, peer review, and computer reminders.

Audit represents a monitoring system on specific aspects of care. It is often a rather formal system, set up and organised by colleges and regional or national committees.[4] The subject of an audit system may vary strongly, and the same applies for the intensity of the related interventions. Feedback resembles audit in many ways, although it tends to be less formal and its development is often dependent on local or even personal initiatives. In peer review, actual performance is reviewed by expert colleagues. Peer review is used not only to improve aspects of clinical care, but also to improve organisational aspects such as practice management. An intervention that needs special attention is the provision of computer reminders. Having the same background and intentions as audit and feedback, these do not generally involve the monitoring of the performance of specific groups or individual doctors. Here the computer may take over, but the intention is still to provide knowledge and give insight into one's own performance. From the viewpoint of the observed clinician, "anonymised" computer reminder systems may appear less threatening:

there are no peers directly involved who must review the actions of those whose behaviour is monitored.

Organisational interventions are focused on facilitating or reducing certain (diagnostic) actions. Examples are changing test ordering forms by restricting the number of tests that can be ordered, and introducing specific prerequisites or criteria that must be met before a test is ordered. There are two structural implementation strategies that can be used to change test ordering. Regulations include interventions where financial incentives or penalties can be easily introduced. Reimbursement systems by health insurance companies or the government may act as a stimulus to urge clinicians to move in the desired direction. Combinations of regulatory steps and financial changes are also conceivable. In several western countries the healthcare system includes payment for tests ordered by doctors, even if the tests are performed by another professional or institution. Adaptation of the healthcare regulations could change this payment system, which might reduce the ordering of too many tests, thereby directly increasing clinicians' income. Even negative incentives for non-rational test ordering can be built in, acting more or less as a financial penalty.

For obvious reasons we should seek and apply methods that both improve the rationality of test ordering and can stop the increase in the use of tests. Ideally, such instruments combine a more rational test ordering behaviour with a reduction in requests. To determine these, a closer look at the performance of the specific implementation strategies is needed. However, not all of these instruments have been used on a regular basis. Some have been regularly applied on a variety of topics (such as changing test ordering), whereas others have been used only incidentally. Most implementation strategies try to change clinicians' behaviour. Although evidence showing relevant effects is required to justify their implementation in practice, because of the nature of the intervention it is not always possible to perform a proper effect evaluation. Especially in large-scale interventions, such as changes in the law or healthcare regulations (nationwide), it is virtually impossible to obtain a concurrent control group. Nevertheless, a number of (randomised) trials have been performed, albeit predominantly on relatively small-scale implementation strategies. The next section gives an overview of the experience so far with most of the aforementioned instruments.

Is a change feasible?

Implementation strategies

As shown above, a variety of implementation strategies is available. However, not all of them are successful: some strategies have proved to be effective, but others have been disappointing. This section gives a review

of the literature to assess which implementation strategies are potentially successful and which are not. There have been a number of published reviews focusing on the effectiveness of implementation strategies. Although their conclusions vary, some general consensus can be observed: there are some implementation strategies that on the whole seem to fail and some that are at least promising.

One of the reasons for the increasing use of tests is the simple fact that tests are both available and accessible. Consequently, a simple strategy would be to reduce the availability of tests on forms, or to request an explicit justification for the test(s) ordered. Such interventions have by and large proved to be effective, with a low input of additional costs and effort: Zaat and Smithuis[5,6] found reductions of 20–50%. A drawback of these interventions, however, is that they risk a possible underuse of tests when the test order form is reduced too extensively and unselectively. Therefore, the changes to the form should be selected and designed very carefully.

Among the implementation strategies that have been used and studied regularly in the past are audit and feedback. To that end, tests ordered are reviewed and discussed by (expert) peers or audit panels. Within this group of interventions there is a huge variation in what is reviewed and discussed, in how often and to whom it is directed, and in the way the review is presented. It may not be surprising that the evaluation of various types of peer review does not allow a uniform conclusion. An intervention with substantial effects, also proven in a randomised trial, was feedback given to individual general practitioners, focusing on the rationality of tests ordered.[7] After 9 years there was a clear and constant reduction in test use, mainly due to a decrease in non-rational requests. Not all interventions are so successful. In studies by Wones[8] and Everett,[9] feedback was restricted to information about the costs of tests ordered; it did not show any effect. Nevertheless, there is evidence suggesting that feedback under specific conditions is an effective method to bring about change. Feedback is more effective when the information provided can be used directly in daily practice, when the doctor is individually addressed, and when the expert peer is generally respected and accepted.

An increasingly popular implementation strategy is the use of computer reminders. This is stimulated by the explosive growth in the use of computer technology in health care, especially in the last decade. The results of computer reminders are promising. It appears to be a potentially effective method requiring relatively little effort. Reminders have significant but variable effects in reducing unnecessary tests, and seem to improve adherence to guidelines.[10] To date, there are relatively few studies on computer reminders. It may be expected, however, that in the near future more interventions on the basis of computer reminders will be performed as a direct spin-off of the growing use of computer-supported communication facilities in health care.

For one implementation strategy in regular use it is clear that the effects on test ordering behaviour are only marginal, if not absent. For many years much effort has been put into postgraduate education courses and in writing papers for clinical journals. The goal of both is to improve the clinical competence of a large target group. Nowadays there is evidence that the direct effects of these methods are disappointing. In a recently published systematic Cochrane Review, Davis et al.[11] concluded that "didactic sessions do not appear to be effective in changing physician performance". In another Cochrane Review by van der Weijden et al.,[12] it was found that the effects of written material on test ordering was small.

Perpetuation and continuation

One aspect that needs more attention in the future is the perpetuation of interventions once they have been started. It is by no means assured that effects, when achieved, will continue when the intervention itself is stopped. In most studies, the effects after stopping the intervention are not monitored. As one of the exceptions, Tierney[13] performed a follow up after ending his intervention through computer reminders on test ordering. The effects of the intervention disappeared 6 months after it was stopped. On the other hand, Winkens[7] found that feedback was still effective after being continued over a 9-year period. This argues in favour of a continuation of an implementation strategy once it is started.

Evaluating the effects

There is a growing awareness that the effects of interventions are by no means guaranteed. Consequently, to discriminate between interventions that are successful and those that are not, we need evidence from scientific evaluations. However, after a series of decades where many scientific evaluations of implementation strategies have been performed and a number of reviews have been published, many questions remain and final conclusions cannot yet be drawn. In a dynamic environment such as the (para)medical profession, it is almost inevitable that the effects of interventions are dynamic and variable over time too. Hence there will always be a need for scientific evaluation.

As is the case with all scientific evaluations, there are quality criteria that studies should meet.[14] Regarding these criteria, evaluation studies on implementation strategies do not essentially differ from other evaluations. The randomised controlled trial still remains the "gold standard". However, there are some circumstances that need special attention.[15] A striking one is the following. In most studies on improving test ordering behaviour, the doctor is the one whose decisions are to be influenced. This automatically means that the unit of randomisation, and hence the unit of

analysis, is the individual doctor. As the number of doctors participating in a study is often limited, this may have a considerably negative effect on the power of the study. A potential solution to this problem may be found in multilevel analyses.[16]

Cost effectiveness of implementation

As far as the cost effectiveness of intervention strategies is concerned, those that combine good effects with the least effort and lowest costs are to be preferred. On the other hand, we may question whether strategies that so far have not proved to be effective should be continued. Should we continue to put much effort into CME, especially in single training courses or lectures? Who should we try to reach through scientific and didactic papers: the clinicians in daily practice, or only the scientist and policy maker with special interest? Should we have to choose the most effective intervention method, regardless of the effort that is needed? If we start an intervention to change test ordering, does this mean it has to be continued for years? There is no general answer to these questions, although the various reviews that have been published argue in favour of combined, tailormade interventions. How such a combination is composed depends on the specific situation (such as local needs and healthcare routines, and the availability of experts and facilities). Generalisable recommendations for specific combinations are therefore not possible or useful. However, if we look at costs in the long term, computer interventions look quite promising.

From scientific evidence to daily practice

An important objective in influencing test ordering behaviour is the change in the rationality and volume of orders, thereby reducing costs or achieving a better (so-called) cost–benefit ratio. However, the ultimate goal is to improve the quality of care for the individual patient. It might be asked to what extent patients are willing to pay for expensive diagnostic activities, weighing the possibility of achieving better health through doing (and paying) the diagnostics, versus the risk of not diagnosing the disease and staying ill (or getting worse) because of not doing so. In other words, how is the cost–utility ratio of diagnostic testing assessed by the patient? In this context the specific positive or negative (side) effects of (not) testing on the health status of the individual patient are difficult to assess independently of other influences. On the other hand, a reduced use in unnecessary, non-rational tests is not likely to cause adverse effects for the individual.

Despite the increasing research evidence showing the need for changes in test ordering behaviour, doctors will always decide on more than merely scientific evidence when the question whether or not to order a test for a certain patient is at stake.[17] Low diagnostic accuracy or high costs of testing may conflict with a patient's explicit wish to have tests ordered, or with the doctor's wish to gain time, the fear of missing an important diagnosis, his or her feeling insecure, and the wish of both patient and doctor to be reassured. These dilemmas are influenced by a variety of doctor and patient related aspects. Regarding the doctor, one could think of the way in which they were trained, how long they have been active in patient care, the number of patients on their list, their relationship with their patients, and their personal experience with "missing" relevant diseases in the past. The patient might suffer from a chronic disease, or from recurrent vague or unexplained complaints, making them question the skills of the doctor. For this latter category of patients in particular, doctors might order more tests than are strictly necessary. Research into the ways of improving test ordering in these situations is urgently needed.

References

1 Grol RPTM, van Everdingen JJE, Casparie AF. *Invoering van richtlijnen en veranderingen*. Utrecht, De Tijdstroom:1994.
2 (a) Geijer RMM, Burgers JS, van der Laan JR, *et al*. *NHG-Standaarden voor de huisarts I*, 2nd edn. Utrecht, Bunge:1999.(b) Thomas S, Geijer RMM, van der Laan JR, *et al*. *NHG-Standaarden voor de huisarts II*. Utrecht, Bunge:1996.
3 Dinant GJ, van Wijk MAM, Janssens HJEM, *et al*. NHG-Standaard Bloedonderzoek. *Huisarts Wet* 1994;**37**:202–11.
4 Smith R. *Audit in action*. London, BMJ Publishers:1992.
5 Zaat JO, van Eijk JT, Bonte HA. Laboratory test form design influences test ordering by general practitioners in the Netherlands. *Med Care* 1992;**30**:189–98.
6 Smithuis LOMJ, van Geldrop WJ, Lucassen PLBJ. Beperking van het laboratorium-onderzoek door een probleemgeorienteerd aanvraagformulier [abstract in English]. *Huisarts Wet* 1994;**37**:464–6.
7 Winkens RAG, Pop P, Grol RPTM, *et al*. Effects of routine individual feedback over nine years on general practitioners' requests for tests. *BMJ* 1996;**312**:490.
8 Wones RG. Failure of low-cost audits with feedback to reduce laboratory test utilization. *Med Care* 1987;**25**:78–82.
9 Everett GD, de Blois CS, Chang PF, Holets T. Effects of cost education, cost audits, and faculty chart review on the use of laboratory services. *Arch Intern Med* 1983;**143**:942–4.
10 Buntinx F, Winkens RAG, Grol RPTM, Knottnerus JA. Influencing diagnostic and preventive performance in ambulatory care by feedback and reminders. A review. *Fam Pract* 1993;**10**:219–28.
11 Davis D, O'Brien MA, Freemantle N, *et al*. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? *JAMA* 1999;**282**:867–74.
12 van der WeijdenT, Wensing M, Grol RPTM, *et al*. Interventions aimed at influencing the use of diagnostic test. The Cochrane Library 2001 (accepted for publication).
13 Tierney WM, Miller ME, McDonald CJ. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *N Engl J Med* 1990;**322**: 1499–504.

14 Pocock SJ. *Clinical trials: a practical approach*. Chichester, John Wiley & Sons: 1991.
15 Winkens RAG, Knottnerus JA, Kester ADM, Grol RPTM, Pop P. Fitting a routine health-care activity into a randomized trial: an experiment possible without informed consent? *J Clin Epidemiol*. 1997;**50**:435–9.
16 Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract* 2000; **17**: 192–6.
17 Knottnerus JA, Dinant GJ. Medicine-based evidence, a prerequisite for evidence-based medicine. *BMJ* 1997;**315**:1109–10.

12 Epilogue: overview of evaluation strategy and challenges

J ANDRÉ KNOTTNERUS

Summary box

- The first phase in the evaluation of diagnostic procedures consists of (1), specifying the clinical problem, the diagnostic procedures(s), and the research question; (2), a systematic search and review of the literature, to decide whether the question can already be answered or whether a new clinical study is necessary.
- In preparing a new clinical study, the investigator must decide about the need for (1), evaluation of test accuracy in circumstances of maximum contrast or, as a further step, in the "indicated" clinical population; (2), evaluation of the impact of the test on clinical decision making or prognosis. The answers to these questions are decisive for the study design.
- Systematic reviews and meta-analysis, clinical decision analysis, cost effectiveness analysis, and expert panels can help to construct and update clinical guidelines.
- Implementation of guidelines should be professionally supported and evaluated, in view of what is known about how clinicians approach diagnostic problems.
- Further developments in three fields are especially important: innovation of (bio)medical knowledge relevant for diagnostic testing; the development of information and communication technology in relation to clinical research and practice; and, further exploration of methodological challenges in research on diagnostic problems.

Introduction

The chapters in this book speak for themselves and there is no need to repeat them or summarise their contents. However, a compact overview of important steps in the evaluation of diagnostic procedures may be useful. In addition, challenges for future work are outlined.

Important steps

The most important steps are represented in a flow diagram of the evaluation strategy (Figure 12.1), with reference to chapters in the book.

The first step is to specify the clinical problem and the diagnostic procedure(s) to be evaluated, and the aim of the study. Are we looking for the (added) diagnostic value of the procedure, the impact of the procedure on clinical management or on the patient's prognosis, or its cost effectiveness? After having formulated the research question accordingly, we should search and systematically review the literature and decide whether sufficient research data are already available. If not, a new clinical study should be considered.

In preparing a clinical study, to choose the appropriate design the following questions need to be answered: (1) is evaluation of accuracy of the test procedure in ideal circumstances of maximum contrast (still) necessary? (2) has this already been successfully achieved, and should accuracy still be established in the "indicated" clinical population? (3) is the impact of the diagnostic procedure on clinical decision making or prognosis yet unknown? The answers to these questions are decisive for the study design, as shown in Figure 12.1. It is sometimes possible to include more than one type of design in one study. For example, test accuracy can sometimes be determined in the context of a randomised trial or a before–after study.

In preparing and reporting the results of a clinical study, the generalisability or external (clinical) validity should be carefully considered. If the study is unique with regard to clinical applicability, the results represent an important evidence base themselves. More often they contribute to a broader knowledge base and can be included in a systematic review or meta-analysis. Clinical decision analysis, cost effectiveness analysis, and expert panels are helpful in constructing or updating clinical guidelines. The implementation of guidelines in clinical practice should be professionally supported and evaluated, in view of the acquired insights into the way clinicians approach diagnostic problems.

Challenges

Throughout this book a comprehensive range ("architecture") of methodological options have been described. At the same time, it has become clear that there are important challenges for future work.

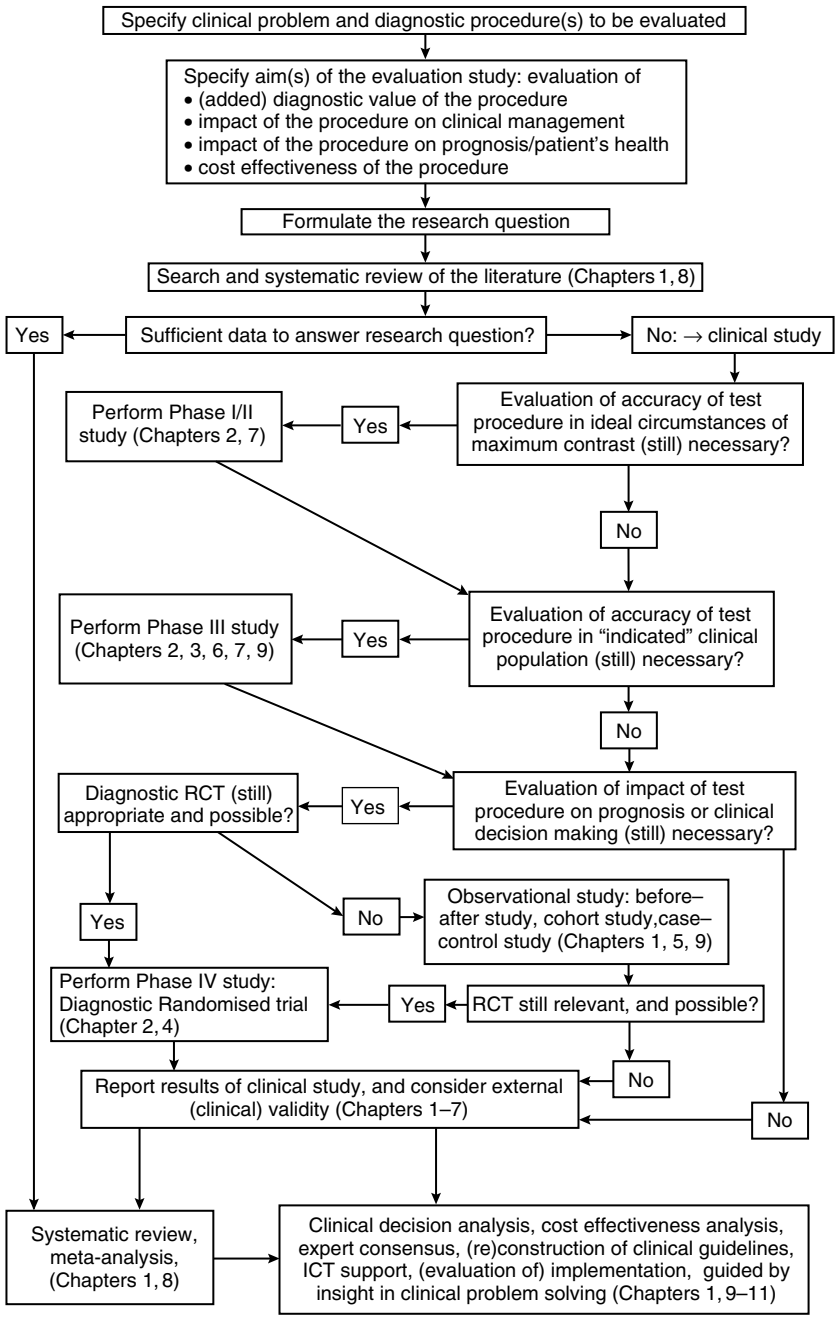


Figure 12.1 Important steps in the evaluation of diagnostic procedures.

Developments in three fields are especially important. First, (bio)medical knowledge will continue to be innovated, cumulated and refined. Second, information and communication technology will further develop and its use in daily practice will be more and more facilitated. Third, in research on diagnostic problems, methodological requirements and limitations are to be increasingly explored.

Innovation of biomedical knowledge and understanding of pathophysiological processes are the principal requirements for the development and evaluation of better diagnostic tests. A clear example is the current work to develop DNA tests in various clinical fields. In future these will not only be supportive of genetic counselling and prenatal screening, but also for clinical diagnostic and prognostic purposes. In addition, the ambition is to use DNA testing to improve the targeting and dosing of therapeutic and preventive interventions ("diagnostic effect modification"). Much work in this field is being done, for example, in cardiovascular medicine,[1] oncology,[2,3] and psychiatry.[4] However, it will be quite some time before these promises will really have an impact on daily patient care. Considerable efforts are still needed, both in the laboratory and in clinical epidemiological research. It was recently shown that the clinical epidemiological quality of many molecular genetic studies was poor: of 40 evaluated research papers, 35 failed to comply with one or more of seven essential methodologic criteria.[5] Furthermore, long-term follow up to clinically validate diagnostic and prognostic predictions needs more attention. In view of the ambition to develop a more tailormade, perhaps even individualised, diagnostic process, population oriented validations will be increasingly unsatisfactory. Also, ethical issues regarding the privacy of genetic information and the right to (not) know have to be dealt with. Doctors and patients, traditionally struggling to reduce diagnostic and prognostic uncertainty, must now learn to cope with approaching certainty.

Although until now computer decision support systems seem to have had more impact on the quality of drug dosing and preventive care than on diagnosis,[6] the growing body and complexity of knowledge enhances the need for (online) diagnostic support systems. The development and evaluation of such systems will therefore remain an important challenge. The same applies to the provision of appropriate input, that is, valid diagnostic and prognostic knowledge. However, performing individual studies in large study populations is very expensive and will always cover only a limited part of diagnostic management. Moreover, such studies may produce results with rather limited generalisability in place and time. Consequently, ways are sought to more efficiently and permanently harvest clinical knowledge and experience. It is therefore worth considering whether and under what conditions accuracy studies, RCTs, quasi-experimental studies, and before–after studies can be more embedded in routine health care.[7] In view of continuity, up to date results, and (external)

clinical validity, much is expected from standardised clinical databases, with international connections, to be used as sampling frames for research. As these databases will be closely related to or even integrated into routine health care, additional efforts are required to meet basic quality and methodological standards and to avoid biases, as discussed in Chapter 10.[8] In the context of such an integrated approach, the implementation of new findings can be studied and monitored in real practice.[9–11]

Diagnostic research should be refined with respect to strategy, spectrum and selection effects, prognostic reference standards,[12] and the assessment of the clinical impact of testing. Data analysis needs progress with regard to "diagnostic effect modification", multiple test and disease outcome categories, and estimation of sample size for multivariate problems. In addition, more flexible modelling is needed to identify alternative but clinically and pathophysiologically equivalent models, appropriate to classify subgroups with varying sets of clinical characteristics with a maximum of predictive power.[13,14] Better methods to improve and evaluate external clinical validity are also required. Furthermore, one must neither forget nor underestimate the diagnostic power of "real life doctors": at least, the performance of proposed diagnostic innovations should be compared with the achievements of experienced clinicians, before they are recommended as bringing new possibilities. We also need more understanding of the "doctor's black box" of diagnostic decision making, using cognitive psychological methods. This can help in more efficient diagnostic reasoning, and in the development of custom-made support systems.[15] Efficiency and speed in the evaluation of the impact of diagnostic procedures can be gained if new data on a specific aspect (for example a diagnostic test) can be inserted into the mosaic of available evidence on a clinical problem, rather than studying the whole problem again whenever one element has changed. For this purpose, flexible scenario models of current clinical knowledge are needed.

Systematic review and meta-analysis of diagnostic studies[16,17] must become a permanent routine activity of professional organisations producing and updating clinical guidelines. Meta-analysis should not only be performed on already reported data, but increasingly also on original and even prospectively developing databases. Such databases can originate from specific (collaborative) research projects, but sometimes also from health care (for example clinics where systematic work ups for patients with similar clinical presentations are routine, and where the population denominator is well defined). Accordingly, meta-analysis, evaluation research, and health care can become more integrated.

The role of the patient in diagnostic management is becoming more active. People want to be involved in the decision as to what diagnostics are performed, and want to know what the outcome means to them. Patient decision support facilities, at the doctor's office and at home, using e-mail

or internet services, are receiving increasing attention. Clinicians have to think about their possible future role in sifting, explaining, and integrating information via these facilities. The question as to which level of certainty is worth which diagnostic procedures, is not always similarly answered by patients and doctors. Patients' perceptions, preferences, and responsibilities should be respected and supported, not excluded, in diagnostic research and guidelines.[18] Sometimes, however, these features are not easily measurable and show substantial inter- and intrasubject variability. A good patient–doctor dialogue therefore remains the core instrument of individual decision making.[19]

Last but not least, formal standards for the evaluation of diagnostics are needed to control acceptance, maintenance, and substitution in the healthcare market. This also requires high quality and transparency of evaluation reports. The initiative to reach international agreement on Standards for Reporting Diagnostic Accuracy (STARD) therefore deserves full support from the scientific and healthcare community.[20]

References

1 Aitkins GP, Christopher PD, Kesteven PJL, *et al*. Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. *Lancet* 1999;**353**:717–19.
2 McLeod HL, Murray GI. Tumour markers of prognosis in colorectal cancer. *Br J Cancer* 1999;**79**:191–203.
3 Midley R, Kerr D. Towards post-genomic investigation of colorectal cancer. *Lancet* 2000; **355**:669–70.
4 Joober R, Benkelfat C, Brisebois K, *et al*. T102C polymorphism in the 5HTA gene and schizophrenia: relation to phenotype and drug response variability. *J Psychiatry Neurosci* 1999;**24**:141–6.
5 Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. *JAMA* 1999;**281**:1919–26.
6 Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA* 1998;**280**:1339–46.
7 van Wijk MA, van Der Lei J, Mosseveld M, Bohnen AM, van Bemmel JH. Assessment of decision support for blood test ordering in primary care. A randomized trial. *Ann Intern Med* 2001;**134**:274–81.
8 Knottnerus JA. The role of electronic patient records in the development of general practice in the Netherlands. *Meth Info Med* 1999;**38**:350–5.
9 Haines A, Rogers S. Integrating research evidence into practice. In: Silagy C, Haines A, eds. *Evidence based practice in primary care*. London: BMJ Books, 1998;pp.144–59.
10 Baker R, Grol R. Evaluating the application of evidence. In: Silagy C, Haines A, eds. *Evidence based practice in primary care*. London: BMJ Books, 1998;pp.75–88.
11 Kidd M, Purves I. Role of information technology. In: Silagy C, Haines A, eds. *Evidence based practice in primary care*. London: BMJ Books, 1998;pp.123–8.
12 Lijmer JG. *Evaluation of diagnostic tests: from accuracy to outcome*. Thesis. Amsterdam: University of Amsterdam, 2001.
13 Heckerling PS, Conant RC, Tape TG, Wigton RS. Reproducibility of predictor variables from a validated clinical rule. *Med Decision Making* 1992;**12**:280–5.
14 Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. *Primary Care* 1995;**22**:341–63.

15 Friedman CP, Elstein AS, Wolf FM, *et al*. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA* 1999;**282**:1851–6.
16 Irwig L, Tosteson AN, Gatsonis C, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.
17 Kester AD, Buntinx F. Meta-analysis of ROC curves. *Med Decision Making* 2000; **20**:430–39.
18 Nease RF, Owens DK. A method for estimating the cost-effectiveness of incorporating patient preferences into practice guidelines. *Med Decision Making* 1994;**14**:382–92.
19 Sackett DL. Evidence-based medicine. *Semin Perinatol* 1997;**21**:3–5.
20 STARD Steering Committee. The STARD project. Standards for Reporting Diagnostic Accuracy. Amsterdam: Department of Clinical Epidemiology, Academic Medical Center Amsterdam, 16–17 December 2000.

Index

Page numbers in **bold** type refer to figures; those in *italic* refer to tables or boxed material