Diagnostic Research

Theory and Application

Carl Moons

ACKNOWLEDGEMENTS

Financial support by the Netherlands Heart Foundation and the Rotterdam Medical Research Foundation (ROMERES) for the publication of this thesis is gratefully acknowledged. The author also gratefully acknowledges the contribution of the Netherlands Institute for Health Sciences (NIHES) to the publication of the thesis.

Cover design: Maria do Sameiro Joia en Eric Neeleman

Printing: Haveka B.V., Alblasserdam.

ISBN 90-9009727-9

Diagnostic Research

Theory and Application

Diagnostisch Onderzoek

Theorie en Toepassing

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam op gezag van de
rector magnificus Prof. Dr P.W.C. Akkermans M.A.
en volgens besluit van het College voor promoties

De openbare verdediging zal plaatsvinden op
woensdag 30 oktober 1996 om 15.45 uur.

door

Karel Gerardus Maria Moons

geboren te Nijmegen

PROMOTIECOMMISSIE:

Promotores:	Prof. Dr. D.E. Grobbee Prof. Dr. J.D.F. Habbema
Overige leden:	Prof. Dr. J. Lubsen Prof. Dr. A.L.M. Verbeek Dr. H.R. Büller
Co-promoter:	Dr. G.A. van Es

*"Theories and ideologies exist in order to escape from the actual.
They prevent seeing what actually takes place, what actually is.
We never question, we just accept them."*

J. Krishnamurti

Ter nagedachtenis aan mijn vader
Voor mijn moeder
Voor Sam

CONTENTS

Manuscripts based on the studies described in this thesis

Chapter 2.1:

Moons KGM, Es GA van, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology, in press.

Chapter 2.2:

Moons KGM, Es GA van, Michel BC, Büller HR, Habbema JDF, Grobbee DE. Hazards of an univariable approach in diagnostic test evaluation. Submitted.

Chapter 2.3:

Moons KGM, Stijnen T, Es GA van, Michel BC, Büller HR, Grobbee DE, Habbema JDF. Treatment thresholds in diagnostic test evaluation: an alternative approach to the comparison of areas under the receiver operating characteristic curve. Med Decis Making, in press.

Chapter 3.1:

Moons KGM, Michel BC, Büller HR, Habbema JDF, Grobbee DE. Evaluation of the independent diagnostic determinants of pulmonary embolism among patient history, physical examination, blood gas analysis, chest X-ray and perfusion lung scintigraphy. Submitted.

Chapter 3.2:

Moons KGM, Klootwijk P, Meij SH, Lenderink T, Baardman T, Es GA van, Habbema JDF, Grobbee DE, Simoons ML. Continuous ST-segment monitoring to predict infarct size and left ventricular function. Submitted.

Chapter 3.3:

Moons KGM, Es GA van, Stijnen T, Bak AAA, Hofman A, Jonker JC, Habbema JDF, Grobbee DE. Efficiency optimization of the selection period in therapeutic trials. J Clin Epidemiol, in press.

CHAPTER 1

INTRODUCTION

VIEWS ON DIAGNOSTIC RESEARCH

> *"Identify the sensitivity and specificity of the sign, symptom, or diagnostic test you plan to use. Many are already published and subspecialists worth their salt ought either to know them from their field or be able to track them down".*[1]

To set a diagnosis in a patient is one of the key challenges in medical practice and forms the basis for clinical care. Diagnosis is not an aim in itself but is relevant in as far as it directs treatment and indicates the prognosis of the patient. Diagnosis amounts to an estimation of the probability of the presence of a particular disease in view of all diagnostic information (patient history, physical examination and test results) in order to decide whether treatment should be initiated or not. A diagnosis is rarely based on one single variable or test and therefore is a multivariable concern per se. However, most diagnostic studies or studies in which diagnostic tests are evaluated still follow a univariable approach. This means that a diagnostic test is evaluated in isolation without explicit regard to the clinical context in which the test is applied. In this respect, clinical practice and diagnostic research frequently do not cohere. In applied medical research of the last decades, little attention has been paid to the principles of diagnostic studies compared to, for example, etiologic studies and studies of treatment efficacy.[2]

Traditionally, diagnostic studies evaluate whether a particular test discriminates between the presence and absence of a particular disease as determined by a reference standard. This research, referred to as *diagnostic accuracy studies*[3-5], is often conducted in a patient population selected on disease status and non-diseased controls. In case of a dichotomous test, the diagnostic accuracy of the test is usually expressed by parameters as the sensitivity, specificity, likelihood ratio (LR), and predictive value (table 1.1). For tests that provide results on a continuous or ordinal scale, the area under the Receiver Operating Characteristic curve (ROC area) is commonly used (figure 1.1). The sensitivity, specificity and to a lesser extent the ROC area, are the most popular measures of diagnostic performance. Because they are conditional on the presence or absence of the disease, whereas diagnosis in practice starts from the presence of symptoms and signs, they have no direct clinical interpretation. Therefore, the use of Bayes' theorem is advocated to estimate diagnostic probabilities. To this aim, sensitivity and specificity or the LR of the applied tests are used together with the prior probability, estimated by the prevalence of disease in the population to which the patient under evaluation belongs.[1,8,9] In this application of Bayes' theorem the user assumes that sensitivity, specificity and LR are constant over patient populations while the prevalence of disease may vary.

Unfortunately, for the same test different diagnostic studies often report different values of the test parameters.

Table 1.1 Characteristics of a dichotomous diagnostic test for a certain disease.

		Disease		
		present	absent	
Test result	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
		TP+FN	FP+TN	

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

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}}$$

$$\text{LR of a positive test result} = \frac{\text{sensitivity}}{1-\text{specificity}} = \frac{\dfrac{\text{TP}}{\text{TP}+\text{FN}}}{\dfrac{\text{FP}}{\text{FP}+\text{TN}}}$$

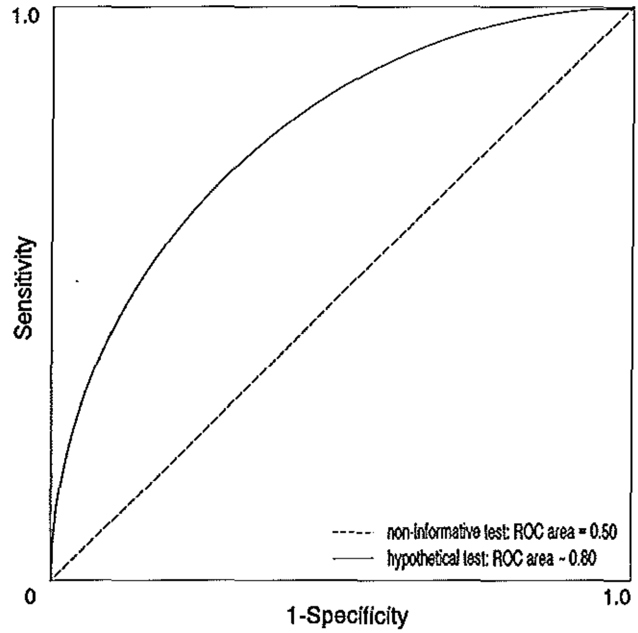
$$\text{LR of a negative test result} = \frac{1-\text{sensitivity}}{\text{specificity}} = \frac{\dfrac{\text{FN}}{\text{TP}+\text{FN}}}{\dfrac{\text{TN}}{\text{FP}+\text{TN}}}$$

$$\text{Predictive value of a positive test result} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Predictive value of a negative test result} = \frac{\text{TN}}{\text{FN}+\text{TN}}$$

+ = positive test result; - = negative test result; TP = true positive diagnosis; FP = false positive diagnosis; FN = false negative diagnosis; TN = true negative diagnosis. TP, FP, FN and FN refer to the observed number of patients in the category.

Figure 1.1 Theoretical example of a ROC curve of a quantitative diagnostic test for a particular disease. A ROC curve plots the sensitivity and 1-specificity as the cut-off value varies over the test result scale. The area under the curve (ROC area) provides an estimate of the overall diagnostic value of the test.[6,7] The diagonal represents a complete non-informative test with a ROC area of 0.50 (for each cut-off value the proportion of true positives equals the proportion of false positives). When the ROC area = 1.0 the test discriminates perfectly between the presence and absence of disease.



This may be due to the use of different cut-off values (in the case of a dichotomized quantitative test), differences in patient selection, selection or verification bias, or a difference in mutual dependencies among the test at issue and other patient characteristics. Many reports have shown that sensitivity and specificity are dependent on the cut-off level that is chosen for test positivity, and that a difference in patient population (e.g. according to symptom or disease severity) may result in different test properties.[10-13] In clinical practice, patients with positive diagnostic test results generally have a higher probability to be referred to further clinical work-up, including a final verification of their disease status, than patients with a negative test result. Accordingly, when diagnostic studies select patients on the true disease status rather than on the patients' indication for diagnostic testing or when diagnostic data obtained from routine clinical practice are used, verification bias may result. This usually leads to biased estimates of the sensitivity and specificity; higher and lower, respectively.[11-15] Other studies have demonstrated that test properties may vary according to clinical (symptoms, signs and other test results) and non-clinical (age, gender and co-morbidities) characteristics of the patient profile, and in particular to the severity of the underlying disease.[10,16,17] This is due to complex mutual dependencies between all diagnostic indicators implying that to some extent different diagnostic indicators provide the same information. Hence, the diagnostic performance of a particular test may (partially) depend on the result of another test. In our view, these mutual dependencies always occur in every patient population. At present, it is generally appreciated that test parameters obtained from a particular population are not directly transferrable to other populations. However, given a certain patient population, diagnostic studies still commonly present *the* sensitivity, *the* specificity, *the* LR and *the* ROC area as a constant and, therefore, as an appropriate indicator of the diagnostic capacity of the test. They are presented as "properties" or "characteristics" of the test for that population. The clinical spectrum of disease manifestations *within* the domain for which the test parameters are estimated is often neglected. Since test parameters seem to vary across subgroups of this clinical spectrum[10,16,17] and *conditional* probabilities of test results given the absence or presence of disease are hardly known[18], the use of these test parameters to estimate diagnostic probabilities in individual patients with Bayes' theorem is questionable. Consequently, if a test is evaluated by the use of "test properties" without consideration of the clinical domain its true diagnostic value remains questionable as well.

The so called "evidence-based medicine working group"[19] and the "outcomes movement"[20], have emphasized the importance of conducting diagnostic research within the relevant clinical setting with use of all routinely obtained data.[1] The aim is to reflect common clinical practice in diagnostic research as a basis for protocol based patient care. In clinical practice, the point of departure is the patient who presents himself with a particular indication for diagnostic testing. A variety of diagnostic tests is often routinely applied. The diagnostic work-up follows a phased approach. Diagnostic indicators from patient history and, subsequently, physical examination are always obtained before the application of diagnostic tests. Subsequent tests may provide additional diagnostic information but may also be burdening for the patient, be time consuming or expensive, and may even produce adverse effects. In agreement with the phased diagnostic work-up in practice, and to consider mutual dependencies, each indicator or test must be evaluated within its diagnostic phase. The fundamental question is whether an indicator or test provides information *added* to the information that is obtained anyway, and in which patient subgroups this is realized. Similarly, when results from other tests are already available, one should question the value of a subsequent test over and beyond the diagnostic information obtained from these previous tests. Thus, the added or independent diagnostic value of a test is important. To conduct diagnostic research in a systematic fashion a database is required with all obtained information on patient history, physical examination, diagnostic tests and final diagnosis. Patients should be selected on their problem of referral (e.g. the indication) to prevent a biased selection on a particular final diagnosis. Using multivariable logistic regression modelling, mutual dependencies and the additional value of each diagnostic indicator can be estimated. In this way, diagnostic probabilities can validly be estimated from all diagnostic determinants simultaneously.[18,21] Diagnostic determinants are those indicators that independently contribute to the prediction of the disease presence. Definition of these determinants should be the objective of diagnostic research; to describe the occurrence or prevalence of a particular disease as a joint function of its diagnostic determinants. Using logistic modelling, the aim should be to construct a diagnostic function which includes those determinants that discriminate well between the absence and presence of the disease and are bearable with respect to patient burden and measurement costs. The trade-off between discriminative power and measurement costs and patient burden is at issue. These considerations of efficiency should motivate diagnostic research but are still largely ignored.

Usually the area under the ROC curve of a diagnostic function is taken to indicate the overall diagnostic value of the function.[22,23] In this way, the diagnostic function represents one (overall) diagnostic test and the individual probability of the presence of disease as estimated by the function represents the "test result". The overall diagnostic value of two or more diagnostic functions can be compared by statistical comparison of their ROC areas.[24] The functions are compared over the entire range of predicted probabilities. However, a decision in the clinical diagnostic work-up commonly reflects a dichotomy (or a trichotomy if additional tests are available). This means that after performing a diagnostic test and adding its information to prior information, the physician has to decide whether to treat or not (or to continue testing). Ideally, the physician wants to bring the previous or prior probability for a patient to have the disease of interest to 1 (absolute certainty on presence) or 0 (absolute certainty on absence). Because this is usually not feasible, the physician endeavours to increase or decrease the prior probability to justify initiation or withholding of treatment, respectively. For example, to initiate treatment the disease must be present with a sufficiently high probability, i.e. the diagnostic probability must exceed a certain threshold. Such threshold is determined by the proportion of misclassifications on disease status (false positive and false negative diagnosis) and the corresponding risks and benefits.[25,26] Physicians intuitively define and apply these thresholds. This suggests that in clinical diagnosis only a specific part of the entire range of "test results" is relevant to decision making. Therefore, the question arises whether the ROC methodology adequately corresponds to clinical practice.

The general aim of this thesis is to outline the principles of clinical diagnostic research and to evaluate methods of diagnostic test evaluation in view of the clinical context. This thesis comprises six studies on diagnostic research. There are three studies (described in chapter 2) that concentrate on the theoretical basis of clinical diagnosis and diagnostic research. Three other studies (described in chapter 3) provide different examples of how the principles of diagnostic research may be applied. Chapter 2.1 discusses the clinical limitations of the conventional diagnostic "test properties". Chapter 2.2 evaluates whether a univariable analysis in diagnostic research or in the evaluation of a diagnostic test in isolation, has relevance from a clinical perspective. Chapter 2.3 examines the use of ROC curves to compare diagnostic tests (or functions) and evaluates an alternative approach to evaluate diagnostic tests taking the (risk and benefits of) subsequent therapeutic decisions into account. Chapter 3.1 evaluates the diagnostic value of patient history, physical examination, and additional tests in patients suspected of pulmonary embolism. Chapter 3.2 evaluates the value of continuous ST-segment monitoring to predict infarct size and left ventricular function in patients with acute myocardial infarction using data from the GUSTO-ischemia monitoring substudy. GUSTO is a large randomized trial to compare four thrombolytic strategies for acute myocardial infarction. Chapter 3.3 describes a study in which the research principles as outlined in previous chapters are applied to increase the efficiency of the selection period of a large primary prevention trial on the efficacy of a cholesterol lowering drug. An approach to improve the cost-effectiveness of the patient selection which amounts to the prediction of eligibility for therapeutic trials, is proposed. In fact, the selection period can be viewed to determine the diagnosis "eligible for the trial". Finally, chapter 4 is a general discussion on the clinical and practical relevance and includes suggestions for further diagnostic research.

References

1. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology; a basic science for clinical medicine. Boston: Little, Brown & Co; 1985.

2. Grobbee DE, Miettinen OS. Clinical epidemiology: introduction to the discipline. Neth J Med 1995;47:2-5.

3. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-94.

4. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071-81.

5. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test: lessons from the rise and fall of dexamethasone suppression test. JAMA 1988;259:1699-1702.

6. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-98.

7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

8. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company; 1985:434-9.

9. Griner PF, Mayewski RJ, Mushlin AL, Greenland P. Selection and interpretation of diagnostic tests and procedures: principles and applications. Ann Intern Med 1981;94:553-600.

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

11. Diamond GA. Selection bias and the evaluation of diagnostic tests: a metadissent. J Chron Dis 1986;39:359-60.

12. Begg CB. Biases in assessment of diagnostic tests. Stat Med 1987;6:411-23.

13. Schouw YT van der, Dijk R van, Verbeek ALM. 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.

14. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.

15. Knottnerus JA, Leffers JP. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epid 1992;45:1143-54.

16. Hlatky MA, Pryor DB, Harell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Am J Med 1984;77:64-71.

17. Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modelling of sensitivity, specificity and predictive value of a diagnostic test. J Clin Epid 1992;45:9-13.

18. Miettinen OS, Caro JJ. Foundations of medical diagnosis: what actually are the parameters involved in Bayes' theorem? Stat Med 1994;13:201-9.

19. Evidence-Based Medicine Working Group. Evidence-based medicine. JAMA 1992;268: 2420-5.

20. Epstein AM. Sounding Board. The outcomes movement-will it get us where we want to go? N Engl J Med 1990;323:266-70.

21. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. New Engl J Med 1985;313:793-9.

22. Bates DW, Cook EF, Goldman L, Lee TH. Predicting bacteraemia in hospitalized patients. A prospectively validated model. Ann Intern Med 1990;113:495-500.

23. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. Ann Intern Med 1990;113:664-70.

24. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.

25. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229-34.

26. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109-17.

CHAPTER 2

THEORY

Chapter 2.1

Limitations of sensitivity, specificity, likelihood ratio and Bayes' theorem in assessing diagnostic probabilities: a clinical example

Introduction

Diagnosis in clinical practice amounts to the estimation of the probability of presence of the target disease given all diagnostic information as obtained from patient history, physical examination and tests results, in an individual patient. To arrive at a diagnosis is one of the key challenges in medical practice, as it forms the basis for clinical work-up; it directs treatment and often is indicative of prognosis. In applied medical research, however, little attention has been paid to the principles of diagnostic research. "Diagnostic" studies usually evaluate whether a particular test in a particular clinical situation can discriminate between presence and absence of disease by calculating sensitivity, specificity and likelihood ratio (LR). These parameters are generally taken as summary measures of performance of diagnostic tests in a certain population although they have no direct diagnostic interpretation. Therefore, the use of Bayes' theorem has been advocated to calculate individual (diagnostic) probabilities by consecutively applying the parameters of all performed tests together with the pre-test probability.[1-6] This application of Bayes' theorem assumes that sensitivity, specificity and LR are constant over patient populations while the pre-test probability may vary. Different studies, however, often report different values of the parameters for the same test. Many reports have demonstrated that test parameters are prone to vary across different patient populations owing to selection or verification bias[6-20], but only few have shown that they may vary across subgroups within a certain population.[21-24]

The aim of this paper is to evaluate the relevance of the sensitivity, specificity and LR of a test in clinical diagnosis, particularly for the same population as that from which the measures are derived. The implications for the use of Bayes' theorem in the assessment of diagnostic probabilities in an individual patient are discussed.

Patients and Methods

Study population

We used cross sectional data from a study conducted in 1988 at the Thoraxcentre, Department of Cardiology, University Hospital Rotterdam, The Netherlands.[25] The study population comprised of 295 subjects consecutively referred by general practitioners to the Thoraxcentre for evaluation of chest pain. All patients had a normal electrocardiography (ECG) at rest, no previous myocardial infarction and did not use digitalis. After informed consent had been given, patient history, physical examination, results from symptom limited exercise testing, and coronary angiography to determine the presence of coronary artery disease (CAD) and the number of diseased vessels, were recorded in that order. Coronary angiography took place within a period of three months after the exercise test and irrespective of its results. A visual reduction of the luminal diameter of at least 50 percent in one or more major arteries at angiography defined the presence of CAD. Two experienced cardiologists who had been blinded to the patient's history and exercise test results independently interpreted the angiograms.

Exercise test

The exercise test was performed in sitting position on a bicycle ergometer, as described earlier.[25] In brief, workload was increased stepwise by 20 Watts per minute until moderate symptoms appeared or exhaustion occurred. Cycling was then continued at a low load for four minutes. Chest electrodes attached at the level of the fifth intercostal space recorded the corrected orthogonal Frank lead ECG. ECG sampling occurred during 20 seconds at rest in the sitting position, and every minute during exercise as well as a six minute recovery period. The sampling frequency was 250 Hz. The baseline level was defined as the mean signal amplitude five to three samples (20-12 msec) before the QRS complex. All amplitudes were measured relative to this baseline. We defined a Heart Rate adjusted ST segment depression (ST/HR) in Frank lead X[26] of 2.0 microvolt/beats per minute or more as a positive, and lower than 2.0 as a negative exercise test. This threshold corresponds approximately to an absolute ST depression of 0.1 millivolt.

$$ST/HR = \frac{(ST_{60}\ \text{at peak exercise} - ST_{60}\ \text{at rest})\ \text{lead}\ X}{\text{heart rate at peak exercise} - \text{heart rate at rest}},$$

where ST_{60} denotes the ST amplitude 60 msec after J-point. We made adjustments for heart rate because many patients were taking beta blocking drugs that were not discontinued during exercise testing and because previous studies have suggested that heart rate adjustment improves the diagnostic potential of the exercise test.[26,27]

Putative determinants of sensitivity, specificity and LR

To study the variation in sensitivity, specificity and LR of the ST/HR depression of exercise testing across patient subgroups, we evaluated determinants of these ST/HR parameters among characteristics of patient history, physical examination, exercise test and underlying disease severity. The patient history and physical examination included age, smoking, diabetes, total cholesterol, systolic blood pressure (baseline SBP), beta blocker use, expected workload (based upon age and height) and symptoms of chest pain. Typical angina was considered to be present if the following three criteria were satisfied: 1) substernal discomfort that was 2) precipitated by exercise, emotion or cold and that was 3) relieved within 10 minutes after rest or sublingual nitroglycerine. We defined "atypical angina" by the presence of two and "non-specific angina" by the presence of only one of the criteria. Beside ST/HR depression, additional variables measured during the exercise test included maximal achieved workload, relative workload (maximal achieved workload/expected workload) and systolic blood pressure at peak exercise. The number of diseased vessels as assessed by coronary angiography defined the categories of severity of disease.

Analyses

Among patients with and without CAD we compared the sensitivity and specificity of the ST/HR response, respectively, across patient subgroups as defined by the above characteristics. We used rate differences and its 95% confidence interval, where the reference group was the subgroup with the lowest sensitivity and specificity. For efficiency, continuous variables were dichotomised to obtain approximately equal numbers without CAD in each subgroup. Based on all subjects, with and without CAD, we calculated the LR (sensitivity/1-specificity) for each patient subgroup. We used the ratio of two likelihood ratio's, where the reference group was the category with the lowest LR, to compare the LR across patient subgroups. Because any two likelihood ratio's are independent (based on two different subgroups), we applied the Taylor series expansion on the two individual standard errors to estimate the standard error of the likelihood ratio ratio.[28] Among patients with and without CAD separately, and following an approach previously proposed by Hlatky et al[21], logistic regression was employed to evaluate which characteristics independently affected exercise test sensitivity and specificity, respectively.[29] We defined the outcome or dependent variable as the positive and negative ST/HR response, respectively. The probability for the outcome, P, can be defined as

$$P = \frac{1}{(1 + \exp^{-(a_0 + b_1.X_1 + b_2.X_2 + \ldots + b_n.X_n)})} ,$$

where a_0 is the intercept and b_1--b_n are the regression coefficients of independent variables X_1--X_n. Given the CAD population, P equals the sensitivity of the ST/HR response corresponding to patient subgroups with observed values of the independent variables. Similarly, given the population without CAD, P is the specificity of the ST/HR response for the subgroup defined by X_1--X_n. We included continuous determinants that had previously been dichotomised as continuous terms in the logistic model if a linear relation was plausible. In the modelling of the sensitivity, we excluded 13 subjects who had missing values. We constructed the multivariable models in concordance with the chronological order in which data generally become available in clinical practice.

Results

207 patients (70 percent) had significant lesions in one or more of the major coronary arteries (table 2.1.1). Of these CAD patients, 119 had a ST/HR depression of 2.0 or higher (sensitivity of 57.5 percent), and 81 of the 88 subjects without CAD had a ST/HR depression lower than 2.0 microvolt/beats per minute (specificity of 92.0 percent). The corresponding LR was 57.5/8.0 = 7.2.

Table 2.1.2 shows that sensitivity of the ST/HR depression substantially differed according to sex, expected workload, absolute achieved workload, relative workload, SBP at peak exercise and number of diseased vessels. Variation over smoking, cholesterol level, baseline SBP and across patients with non-specific and atypical angina, compared with typical angina, was less marked. The specificity differed according to sex, diabetes, baseline SBP and relative workload. Although sensitivity and specificity were conversely affected by most variables, the LR of the exercise test still varied over categories of sex, smoking, cholesterol level, baseline SBP, relative workload and SBP at peak exercise.

Among all variables of patient history and physical examination, sex, baseline SBP, expected workload and cholesterol level were independent determinants of the

Table 2.1.1 Exercise test response, characterised by the dichotomised heart rate adjusted ST (ST/HR) depression, of patients with and without CAD.

ST/HR depression (microvolt/bpm)	CAD patients (n = 207)	non-CAD patients (n = 88)	LR
≥ 2.0	119 (57.5%)	7 (8.0%)	7.2
< 2.0	88 (42.5%)	81 (92.0%)	
Total	207	88	

bpm, beats per minute; LR, likelihood ratio

ST/HR sensitivity (table 2.1.3). Addition of all exercise test variables to the previous model showed that relative workload and SBP at peak exercise were also independent determinants. Since 60 percent of the patients with CAD had a relative workload lower than 75 and 25 percent higher than 95, we included relative workload as a dichotomous rather than as a continuous variable. When we added disease specifications to the previous model, sex, SBP at peak exercise, relative workload and multivessel disease remained as the strongest independent determinants of exercise test sensitivity. From all characteristics of patient history and physical examination, sex was the most important determinant of exercise test specificity (table 2.1.3). After including all exercise test variables to this model, relative workload appeared to be the only independent determinant of the specificity. The odds ratio for sex was very imprecise. Using thresholds of ST/HR depression other than 2.0, the sensitivity, specificity and LR of the exercise test similarly varied according to patient characteristics. We also found similar results when we used other outcome parameters of the exercise test, such as absolute ST depression at maximal workload or ST depression at maximal workload relative to rest (data not shown). These results agree with previous studies.[21,30-32]

Discussion

This study demonstrates that sensitivity, specificity and LR of the ST/HR depression of exercise testing are not constant but vary across subgroups as defined by various

Table 2.1.2 Variations in sensitivity, specificity and LR of the heart rate adjusted ST depression according by various characteristics of patients with and without CAD, expressed as rate difference and likelihood ratio ratio.

Patient characteristic	CAD patients			non-CAD patients				
	n	Sens (%)	Rate difference (95% CI)	n	Spec (%)	Rate difference (95% CI)	LR	Likelihood ratio ratio (95% CI)
History + Physical								
Age (years)								
28-50	71	57.8	0.4 (-13.8 to	46	91.3	-	6.6	-
51-70	136	57.4	-	42	92.9	1.6 (- 9.7 to	8.1	1.2 (0.3- 5.2)
Sex								
male	170	63.5	33.8 (17.4 to	52	88.5	-	5.5	-
female	37	29.7	-	36	97.2	8.7 (-1.4 to	10.7	1.9 (0.2-16.4)
Symptoms								
non-specific	18	50.0	-	37	91.9	3.7 (-14.0 to	6.2	1.2 (0.2- 6.9)
atypical	55	50.9	0.9 (-25.7 to	34	94.1	5.9 (-11.4 to	8.7	1.7 (0.3-11.1)
typical	134	61.2	11.2 (-13.3 to	17	88.2	-	5.2	-
Diabetes¶								
yes	29	62.1	6.3 (-12.9 to	5	100	9.1 (0.1 to	∞	‖
no	172	55.8	-	77	90.9	-	6.1	
Smoking¶								
yes	114	60.5	8.8 (- 5.0 to	31	93.6	3.6 (- 8.4 to	9.4	1.8 (0.4- 9.0)
no	87	51.7	-	50	90.0	-	5.2	-
Beta-blocker use								
yes	123	57.7	0.6 (-13.1 to	45	91.1	-	6.5	-
no	84	57.1	-	43	93.0	1.9 (- 9.3 to	8.2	1.3 (0.3- 5.4)
Cholesterol (mmol/l)¶								
4.0- 6.0	52	51.9	-	33	87.9	-	4.3	-
6.1-12.0	150	61.3	9.4 (- 6.2 to 25.1)	48	93.8	5.9 (- 7.2 to 18.9)	9.9	2.3 (0.5- 9.9)
Expected load (Watt)¶								
70-149	94	50.0	-	45	93.8	3.8 (- 7.8 to 15.3)	8.1	1.3 (0.3- 5.3)
150-240	112	64.3	14.3 (0.8 to 27.7)‖	43	90.0	-	6.4	-

... To be continued

Table 2.1.2 Continued.

Patient characteristic	CAD patients			non-CAD patients				
	n	Sens (%)	Rate difference (95% CI)	n	Spec (%)	Rate difference (95% CI)	LR	Likelihood ratio ratio (95% CI)
SBP* baseline (mmHg)								
100-140	79	64.6	11.4 (-2.2 to 25.1)	52	96.2	10.1 (-2.4 to 22.5)‡	17.0	4.5 (0.9-21.7)
141-240	128	53.1	-	36	86.1	-	3.8	-
Additional test variables								
Maximal load (Watt)								
45-134	162	62.4	22.4 (6.2 to 38.9)†	45	91.1	-	7.0	1.2 (0.3- 5.4)
135-280	45	40.0	-	43	93.0	1.9 (-9.4 to 13.2)	5.7	-
Relative load (%)¶								
30- 90	154	66.2	33.5 (18.8 to 48.3)†	41	85.4	-	4.5	-
91-140	52	32.7	-	47	97.9	12.5 (0.9 to 24.1)‡	15.6	3.5 (0.4-28.1)
SBP peak (mmHg)¶								
110-175	93	66.7	16.7 (3.1 to 30.2)†	42	92.9	2.0 (-9.8 to 13.5)	9.4	1.7 (0.4- 7.3)
176-240	106	50.0	-	44	90.9	-	5.5	-
Disease specifications								
Number diseased vessels								
none				88	90.9			
one	71	39.4	-					
two	74	58.1	18.7 (2.7 to 34.7)†					
three	62	77.4	38.0 (22.6 to 53.4)†					

Sens, sensitivity; Spec, specificity; CI, confidence interval; LR, likelihood ratio; SBP, systolic blood pressure; - , reference category

† Determinant of sensitivity

‡ Determinant of specificity

§ Exact 95% CI of the odds ratio (95% CI of the RD could not be assessed because there were 0 observations in one cell)

‖ Methodology could not be applied to an infinite likelihood ratio

¶ A few values were missing

Table 2.1.3 Logistic regression coefficients and odds ratios for the independent determinants of sensitivity and specificity of the heart rate adjusted ST depression for the three models comprising patient history + physical examination, patient history + physical examination, additional exercise test results, and additional disease specifications.

	Patient history					
	Sensitivity (n=202)			Specificity (n=88)		
Determinant	ß	OR	95% CI	ß	OR	95% CI
Intercept	1.43			3.56		
Male	2.03	7.61	(2.48-23.34)	-1.69	0.18	(0.02-1.42)
Baseline SBP (mmHg)	-0.015	0.99	(0.97- 1.00)	–	–	–
Expected workload (Watt)	-0.013	0.99	(0.97- 1.00)	–	–	–
Cholesterol level (mmol/l)	0.21	1.20	(1.01- 1.50)	–	–	–
SBP peak exercise (mmHg)						
Relative workload 30%-90%						
Multi vessel disease†						

... To be continued

Table 2.1.3 Continued.

	Patient history + physical examination + exercise test						Patient history + physical examination + exercise test + disease specifications		
	Sensitivity (n=194)			Specificity (n=88)			Sensitivity (n=194)		
Determinant	ß	OR	95% CI	ß	OR	95% CI	ß	OR	95% CI
Intercept	1.61			3.83			1.34		
Male	2.03	7.61	(2.35-24.68)	–	–	–	1.10	3.00	(1.17-2.82)
Baseline SBP (mmHg)	-0.011	0.99	(0.97- 1.01)	–	–	–	-0.006	0.99	(0.97-1.01)
Expected workload (Watt)	-0.012	0.99	(0.98- 1.00)	–	–	–	-0.009	0.99	(0.98-1.01)
Cholesterol level (mmol/l)	0.18	1.20	(0.97- 1.49)	–	–	–	0.12	1.13	(0.91-1.03)
SBP peak exercise (mmHg)	-0.018	0.98	(0.97- 0.99)	–	–	–	-0.018	0.98	(0.97-0.99)
Relative workload 30%-90%	1.03	2.80	(1.30- 6.02)	-2.07	0.13	(0.02-1.07)	0.97	2.64	(1.25-5.56)
Multi vessel disease†							0.94	2.56	(1.31-4.98)

ß, regression coefficient; OR, odds ratio; CI, confidence interval; SBP, systolic blood pressure; – , variable was not an independent determinant

† Coded as 1 for two or three diseased vessels to 0 for one diseased vessel

characteristics of the patient profile. Logistic models including the independent determinants of sensitivity and specificity can be used to estimate these parameters for a specific patient subgroup. Given the second models of table 2.1.3, for example, the sensitivity, specificity and LR in men with expected workload 180 Watt, cholesterol level 8.0 mmol/l, SBP at baseline 70 mmHg, SBP at peak exercise 150 mmHg, and relative workload 90 percent or lower would be 61.8 percent, 85.6 percent and 4.3, respectively, whereas in women with these same values the parameters would be 31.4 percent, 97.9 percent and 15.0, respectively.

For some determinants there were missing values. However, the number of missing values was very low and they were rather equally distributed across thè positive and negative exercise test responders. Also, there were no reasons to believe that they were selected from a particular category of the determinant. Therefore, we believe that these missing values have not affected the results.

Because sensitivity and specificity usually vary in opposite directions, the LR is thought to be more stable than sensitivity and specificity. This study showed that the LR similarly differs across patient subgroups of a certain population. Owing to the small number of patients without CAD and a positive test result, we cannot draw a definite conclusion regarding variability of specificity and likelihood ratio, although these appear as unstable as sensitivity for our data. Patient history, physical examination, test results and disease severity are complex and mutually dependent factors that collectively determine sensitivity, specificity and LR of the exercise test.[33] Many patient characteristics, e.g. sex and cholesterol level, are associated with the development, presence and severity of CAD. The severity of CAD determines the probability of a finding a positive (or negative) exercise test result. Hence, patient characteristics also determine the sensitivity (or specificity) of exercise testing, although this becomes less relevant if adjusted for disease severity. Because in clinical diagnosis we never know the underlying disease severity, knowledge of the determinants that can be measured is important. A single level of test parameters for the exercise test that applies to all patient subgroups of the population cannot be found and should not be sought.

Several previous reports have demonstrated variability of sensitivity and specificity across different patient populations. This finding has been related to "selection bias" or "spectrum bias", i.e. selective referral by characteristics (symptomatology or test results) previously documented in the patients.[6-20] It is generally well appreciated that test parameters are not directly applicable to other populations, because patient populations differ in the spectrum of disease manifestations, and sensitivity and specificity vary for different groups across this spectrum.[21-24,34,35] Nevertheless, given a certain population sensitivity, specificity and the LR are usually still presented as pertaining to that domain regardless of the clinical spectrum within that domain.[36,37] Our findings show that even *within* a certain patient population the exercise test parameters may vary substantially across specific categories according to the characteristics of that category. As Diamond commented before[36,38], this variability could still reflect selective referral of positive test responders to disease verification (angiography), although the variations across disease specifications remain hard to explain. Diamond and others showed that just as the predictive value of a test varies owing to variations in the prevalence of disease, test parameters vary across patient subgroups due to variations in the overall frequency of positive test responses in these patient subgroups, i.e. the frequency in all patients who underwent the test.[9,14,19,36,38,39] They proposed ways to adjust for selective referral by positive test results given the overall frequency of positive test responses per patient subgroup. After such adjustment the remaining variation in sensitivity and specificity would partly be reduced to a predictable pattern of variation.[38] In our view, however, the fact remains that the sensitivity, specificity and LR vary across patient subgroups, whether or not this variation is due to the variation in overall test responses. Instead of evaluating determinants of sensitivity and specificity in a certain verified sample as we have done in our study, one could as well study determinants of the overall frequency of a certain test result in the total.tested population to conclude that there are substantial variations in the test parameters across patient subgroups. Note that there is a true sensitivity, specificity and LR for each homogeneous subgroup. Patient populations, however, are always heterogeneous with respect to diagnostic characteristics. We have shown that these characteristics are mutually dependent issues. Hence, a proper definition of these homogene subgroups is difficult if not impossible, even within a certain selected population. As heterogeneity of patient populations pertains to most diseases stability of diagnostic test parameters across that population can generally not be assumed.

Diagnosis is rarely based on one single characteristic or test whereas all involved characteristics are potentially correlated. For a proper interpretation of a particular test all other modifying factors should be considered in clinical diagnosis. Although this is indeed what happens in the mind of the physician, the mutual dependencies may be very complex and hardly distinguishable. In the application of Bayes' theorem to estimate individual probabilities from more than one diagnostic factor, each factor or determinant is considered as a different diagnostic "test" with its sensitivity, specificity and LR. This study, however, has demonstrated that a single level of these parameters does not exist in a common diagnostic setting because of mutual dependencies. Different patient subgroups may have different parameters per diagnostic determinant and the number of these subgroups can be very high. To arrive at a diagnosis in an individual patient using Bayes' theorem the physician needs to know the test parameters of the corresponding patient subgroup, which are hardly ever known. Therefore, the application of Bayes' theorem to estimate diagnostic probabilities for several diagnostic determinants simultaneously has serious limitations. Such estimation is nevertheless possible using a multivariable prevalence function as derived by logistic regression analysis.[40-43] This equation can estimate the probability of the presence of disease given other relevant diagnostic determinants that could possibly modify this probability. This logistic regression model makes no use of sensitivity, specificity or LR. Therefore, these concepts are not required for diagnosis. It should be appreciated that this does not omit Bayesian thinking in diagnosis. A basic principle of clinical practice is that the interpretation of new information depends on a priori beliefs. Diagnosis, accordingly, is a consecutive decision making process. Information from each diagnostic test is added to the prior information in order to decide whether to initiate treatment or to prolonge testing. This decision should be guided by careful judgement of diagnostic probabilities preferably estimated by multivariable logistic regression models instead of using Bayes' theorem. Various models may be constructed and extended in accordance with the diagnostic work-up in practice.

In conclusion, single values for sensitivity, specificity and LR of a test do not exist, and therefore the commonly proposed use of Bayes' theorem has major limitations in the assessment of diagnostic probabilities. Test parameters, however, are still extensively published in the literature as characteristics (properties) of the discriminative power of a test to subscribe it's diagnostic relevance, and are advocated for use in clinical diagnosis. The use of prevalence functions provides an alternative that lacks the limitations inherent to conventional test parameters. This indicates a valuable approach in diagnostic research.

References

1. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130:9-21.

2. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: WB Saunders Company; 1980:92-108.

3. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company; 1985:434-9.

4. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology; a basic science for clinical medicine. Boston: Little, Brown & Co; 1985:110-25.

5. Griner PF, Mayewski RJ, Mushlin AL, Greenland P. Selection and interpretation of diagnostic tests and procedures: principles and applications. Ann Intern Med 1981;94:553-600.

6. Sox HC Jr. Probability theory in the use of diagnostic tests. Ann Intern Med 1986;104:60-6.

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

8. Harris Jr JM. Hazards of bedside Bayes. JAMA 1981;246:2602-5.

9. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.

10. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. N Engl J Med 1983;309:518-22.

11. Rozanski A, Diamond GA, Forrester JS, Berman DS, Morris D, Swan HJC. Alternative referent standards for cardiac normality. Ann Intern Med 1984;101:164-71.

12. Begg CB. Statistical methods in medical diagnosis. CRC Crit Rev Med Inform 1986;1:1-22.

13. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias; application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. J Chron Dis 1986;39:343-55.

14. Diamond GA. Selection bias and the evaluation of diagnostic tests: a metadissent. J Chron Dis 1986;39:359-60.

15. Begg CB, Greenes AR, Iglewicz B. The influence of uninterpretable test results on the assessment of diagnostic tests. J Chron Dis 1986;39:575-84.

16. Begg CB. Biases in assessment of diagnostic tests. Stat Med 1987;6:411-23.

17. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. Med Decis Making 1987;7:115-9.

18. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and disease. Med Decis Making 1987;7:139-48.

19. Diamond GA. Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? Med Decis Making 1991;11;48-56.

20. Knottnerus JA, Leffers JP. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epid 1992;45:1143-54.

21. Hlatky MA, Pryor DB, Harell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Am J Med 1984;77:64-71.

22. Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher VF. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med 1988;84:699-710.

23. Levy D, Labib SB, Anderson KM, Christiansen JC, Kanell WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990;81:815-20.

24. Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modelling of sensitivity, specificity, and predictive value of a diagnostic test. J Clin Epidemiol 1992;45:1-7.

25. Deckers JW, Rensing BJ, Tijssen JGP, Vinke RV, Azar AJ, Simoons ML. A comparison of methods of analysing exercise tests for diagnosis of coronary artery disease. Br Heart J 1989;62(6):438-44.

26. Detrano R, Salcedo E, Passalaqua BA, Friis R. Exercise electrocardiographic variables: a critical appraisal. JACC 1986;8:36-47.

27. Elamin MS, Mary DASG, Smith DR, Liden RJ. Prediction of severity of coronary artery disease using slope of submaximal ST-segment/heart rate relationship. Cardiovasc Res 1980;14:681-91.

28. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epid 1991;44:763-70.

29. Harrel FE. The LOGIST procedure. In: Hastings RP, ed. SUGI supplemental library user's guide. Version 5 edition. Cary, NC: SAS Institute Inc.; 1986:269-93.

30. Rifkin RD, Hood Jr WB. Bayesian analysis of electrocardiographic exercise stress testing. N Engl J Med 1977;297:681-6.

31. Weiner DA, Ryan TJ, McCabe CH, et al. Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). N Engl J Med. 1979;301:230-5.

32. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. Am J Cardiol 1980;46:807-12.

33. Hilden J. Optimistic bias in the assessment of sensitivity and specificity. J Chron Dis 1986;10:853-5.

34. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986;104:66-73.

35. 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;113:147-54.

36. Diamond GA. Clinical epistemology of sensitivity and specificity. J Clin Epid 1992;45:9-13.

37. Feinstein AR. Clinical judgement revisited: the distraction of quantitative models. Ann Intern Med 1994;120:799-805.

38. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. Am J Cardiol 1986;57:1175-80.

39. Greenes RA, Begg CB. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. Invest Radiol 1985;20:751-6.

40. Schmitz PIM, Habbema JDF, Hermans J. The performance of logistic discrimination on myocardial infarction data in comparison with some other discriminant analysis methods. Stat Med 1983;2:199-205.

41. Pryor DB, Shaw L, McCants CB, et al. Value of the history and physical in identifying patients at increased risk for coronary artery disease. Ann Intern Med 1993;118:81-90.

42. Gray D, Hampton JR, Shaw LK, Bernstein SJ, Pryor DB. Succesful application of a predictive model of coronary artery disease [abstract]. Circulation 1992;86:I-41.

43. Miettinen OS, Caro JJ. Foundations of medical diagnosis: what actually are the parameters involved in Bayes' theorem? Stat Med 1994;13:201-9.

Chapter 2.2

Hazards of an univariable approach in diagnostic test evaluation

Introduction

Diagnostic research commonly evaluates whether a test discriminates between the presence and absence of a particular disease, as determined by a reference standard. This research approach has been referred to as 'diagnostic accuracy' or 'test evaluation' studies.[1-8] These studies tend to be conducted in a patient population selected for their disease status rather than on the clinical problem and usually follow a univariable approach. The latter implies that the test is evaluated in isolation without reference to the clinical context in which the test is applied and its results interpreted. Data obtained from patient history, physical examination and other diagnostic tests before the test at issue would be applied, are neglected. Hence, the true clinical value of the diagnostic information provided by the test may be less since the other available data may already provide such information.[6] The diagnostic accuracy of the (single) test is usually expressed by measures such as sensitivity, specificity, likelihood ratio (LR) and the area under the receiver operating characteristic (ROC) curve. These measures are in this paper referred to as test parameters. In practice, diagnosis amounts to an estimation of the probability of the presence of disease given all diagnostic variables (symptoms, signs and test results) documented in a patient presenting with a particular clinical problem. A test is almost invariably part of a set of diagnostic variables. A diagnosis is rarely based on one single variable or test and therefore is a multivariable concern per se. In this respect, clinical practice and diagnostic research frequently do not cohere. The question rises whether diagnostic accuracy studies, when they follow a univariable approach, provide meaningful information in diagnostic test evaluation.

This paper addresses the limitations of an univariable approach in the evaluation of diagnostic tests which is illustrated by data from a study on the diagnosis of pulmonary embolism.

Patients and methods

Patients

Data were used from consecutive patients admitted with clinically suspected pulmonary embolism to the Academic Medical Centre and the Slotervaart Hospital in Amsterdam, The Netherlands.[9-12] Because pulmonary embolism is a life threatening disease, a proper diagnosis in these patients is vital. To this aim, patient history (e.g. age, dyspnoea, gender, previous deep venous thrombosis), physical examination (e.g. wheezing, pleural rub, body temperature), blood gas analysis (arterial oxygen pressure, PaO_2), chest X-ray, leg ultrasonography (to detect the presence of deep leg venous thrombosis), and perfusion and ventilation lung scan results were subsequently obtained. Pulmonary embolism was considered absent (no treatment initiated) or present (treatment initiated) in case of a normal or high probability result of the ventilation-perfusion lung scan, respectively. In 186 of the 452 referred patients with an intermediate non-high probability scan result, pulmonary angiography was used to determine the presence of pulmonary embolism. In 40 patients pulmonary angiography could not be performed because of medical reasons such as manifest heart failure, severe pulmonary hypertension or poor clinical conditions. In another six patients the angiogram was non-interpretable. Data on the remaining 140 patients, of which 38 had angiographically proven pulmonary embolism (prevalence = 28 percent), were used in the current analysis. Pulmonary angiograms were evaluated without knowledge of the other diagnostic information.

The X-thorax was considered abnormal, i.e. possible presence of pulmonary embolism, if it showed a raised diaphragm, pleural effusion, atelectasis, consolidation or signs of heart failure. The leg ultrasound was considered abnormal, i.e. deep leg venous thrombosis was considered present, if the femoral vein and/or popliteal vein were noncompressible.

Data analysis

Data analysis was performed with standard software packages (SAS Institute Inc., Cary release 6-10). The aim of the present study was to compare the results of an univariable evaluation of diagnostic tests with multivariable evaluation of the tests in their appropriate work-up phase in the diagnostic process. For each diagnostic variable we first estimated the sensitivity, specificity, LR of a "positive" [sensitivity/(1-specificity)] and "negative" [(1-sensitivity)/specificity] test result, and the positive and negative predictive value which, defined as the presence of pulmonary embolism given a "positive" and "negative" testresult respectively. To this aim, continuous or categorical variables were dichotomised at a clinically relevant cut-off value. The values of the variable that were indicative for the presence of pulmonary embolism were defined as the "positive result" as opposed to the "negative result". A LR of the positive (LR+) and negative (LR-) testresult equal to unity indicates that the particular value of the variable does not discriminate between presence and absence of pulmonary embolism, whereas a high LR+ and low LR- indicates its presence and absence respectively. For continuous indicators the area under the ROC curve (ROC area) was calculated using a non-parametric approach.[13] The association of each indicator with the presence of disease was estimated by logistic modelling. Continuous variables were entered into the model without categorisation if a linear relation was plausible.

In a previous study we have described the derivation of a clinical decision rule for the diagnosis of pulmonary embolism in non-high probability patients using findings on perfusion lung scintigraphy, history and physical examination.[12] From this analysis, the five independent determinants for the diagnosis of pulmonary embolism were the presence of multiple defects on perfusion lung scanning, new or recently worsened cough, previous deep venous thrombosis, body temperature above 37° Celsius, and the absence of wheezing. Using multivariable logistic regression analysis, and in concordance with the sequence of the diagnostic work-up in practice, we separately included PaO_2, chest X-ray and leg ultrasonography to the model including the five independent predictors obtained from the perfusion lung scan, patient history and physical examination.[12] The objective was to evaluate their independent or added value in the diagnosis of pulmonary embolism in patients with a non-high probability ventilation/perfusion lung scan result. Subsequently, combinations of the three diagnostic procedures added to the initial model were evaluated. This again conforms with the usual diagnostic work-up (blood parameters and chest X-ray are generally obtained prior to leg ultrasound). The ROC area and its standard error[13] were used to compare the diagnostic information content or the discriminative value of all models. The correlation between the models was taken into account because they were based on the same cases.[14] In the multivariable analyses, 30 subjects who had missing values were excluded.

Several authors have suggested to evaluate (differences in) diagnostic test performance across clinically different subgroups.[15-17] Therefore, we applied the diagnostic models (regarded as an overall test) to patient subsets as defined by the number of defects on the initial perfusion lung scan (multiple or single defects). Per subset, the mean predicted probability of the different diagnostic models was compared. This analysis provided a kind of validation study to evaluate the average performance of the models in different patient subsets.

Results

Table 2.2.1 shows the results of univariable analysis. Given a prevalence or prior probability of 0.28, from all history and physical findings the predictive value was relatively high for the presence of previous deep venous thrombosis (0.45), pleural rub (0.41) and a new or recently worsened cough (0.39). Their positive likelihood ratio was also relatively high (2.2, 1.8 and 1.7, respectively) although the corresponding negative likelihood ratio was close to one. These likelihood ratios were associated with very different sensitivities (13, 29 and 45 percent, respectively) and specificities (94, 84 and 74 percent, respectively). In view of the prevalence of 0.28, the negative predictive value was relatively low for multiple perfusion defects (0.17), heart frequency of 95 or higher (0.19) and absence of wheezing (0.07). The latter also had a low likelihood ratio of the negative result, i.e. presence of wheezing. For all other history and physical findings both predictive values were not markedly higher or lower than the prevalence, and both likelihood ratios were close to one. Here too, the likelihood ratios were associated with very different sensitivities and specificities.

The positive predictive value, positive likelihood ratio and specificity (although the latter was based on only four patients) of leg ultrasound were markedly higher compared to these same parameters of chest X-ray and PaO_2 (table 2.2.1). However, chest X-ray had a much higher sensitivity compared to the other two tests and a much lower negative predictive value (0.19) compared to PaO_2 (0.27) but not compared to leg ultrasound (0.21). For any of the continuous variables the ROC area was low, and close to a completely non-informative test with a ROC area of 0.50. Using univariable logistic modelling, cough, wheezing, body temperature, PaO_2, X-thorax and ultrasound were all significantly associated with pulmonary embolism (p-value < 0.05).

In multivariable analysis, the ROC area (figure 2.2.1) of the diagnostic model including the independent determinants of perfusion lung scintigraphy, patient history and physical examination was 0.79 (table 2.2.2). Excluding variables from this model significantly decreased the ROC area. The ROC area marginally increased to 0.81 after addition of PaO_2 (figure 2.2.1), and significantly increased after addition of chest X-ray and leg ultrasound to 0.84 and 0.83, respectively (figure 2.2.2). The diagnostic

Table 2.2.1 Univariable association of diagnostic variables and pulmonary embolism, expressed as sensitivity, specificity, likelihood ratio of positive (LR+) and negative (LR-) result, area under receiver operating characteristic (ROC) curve (for continuous indicators), positive predictive value (PV+, probability of presence of disease given a positive result) and negative predictive value (PV-, probability of presence of disease given a negative result).

	PE (n=38)		No PE (n=102)						
Diagnostic variables	n	Sens (%)	n	Spec (%)	LR+	LR-	ROC area	PV+	PV-
Perfusion lung scan									
Multiple defects on perfusion scan	30	83	73	28	1.2	0.6		0.29	0.17
Segmental defects on perfusion scan	22	61	57	44	1.1	0.9		0.28	0.24
Patient history+physical examination									
Age > 70 years	12	32	34	67	1.0	1.0	0.50	0.26	0.28
Gender (% male)	20	53	48	47	1.0	1.0		0.29	0.25
Cough, new or recently worsened	17	45	27	74	1.7	0.7		0.39	0.22
Previous DVT	5	13	6	94	2.2	0.9		0.45	0.26
Malignancy	8	21	28	73	0.8	1.1		0.22	0.29
Days of immobilization > 1 day†	8	22	32	68	0.7	1.1	0.55	0.20	0.29
Absence of wheezing	36	95	77	25	1.3	0.2		0.32	0.07
Body temperature > 37° Celsius	23	61	41	60	1.0	1.0		0.36	0.33
Pleural rub	11	29	16	84	1.8	0.8		0.41	0.24
Signs of DVT	4	11	10	90	1.1	1.0		0.29	0.27
Respiratory frequency ≥ 20 breaths/min†	21	60	49	51	1.2	0.8	0.58	0.30	0.22
Heart frequency ≥ 95 beats/min†	22	67	52	48	1.3	0.7	0.52	0.30	0.19
Additional tests									
PaO_2 ≥ 80 mmHg†	13	37	24	71	1.3	0.9	0.59	0.35	0.27
Abnormal chest X-ray	34	89	68	67	2.7	0.2		0.33	0.19
Abnormal leg ultrasound†	10	29	4	96	7.3	0.7		0.71	0.21

PE, pulmonary embolism; n, number of patients with the diagnostic test result; sens, sensitivity; Spec, specificity; DVT, deep venous thrombosis; min, minute; PaO_2, arterial oxygen pressure.
† A few values were missing.

Table 2.2.2 Results of the multivariable logistic regression analysis for three different diagnostic models to assess the presence of pulmonary embolism in 110 patients with a non-high probability ventilation-perfusion scan result.

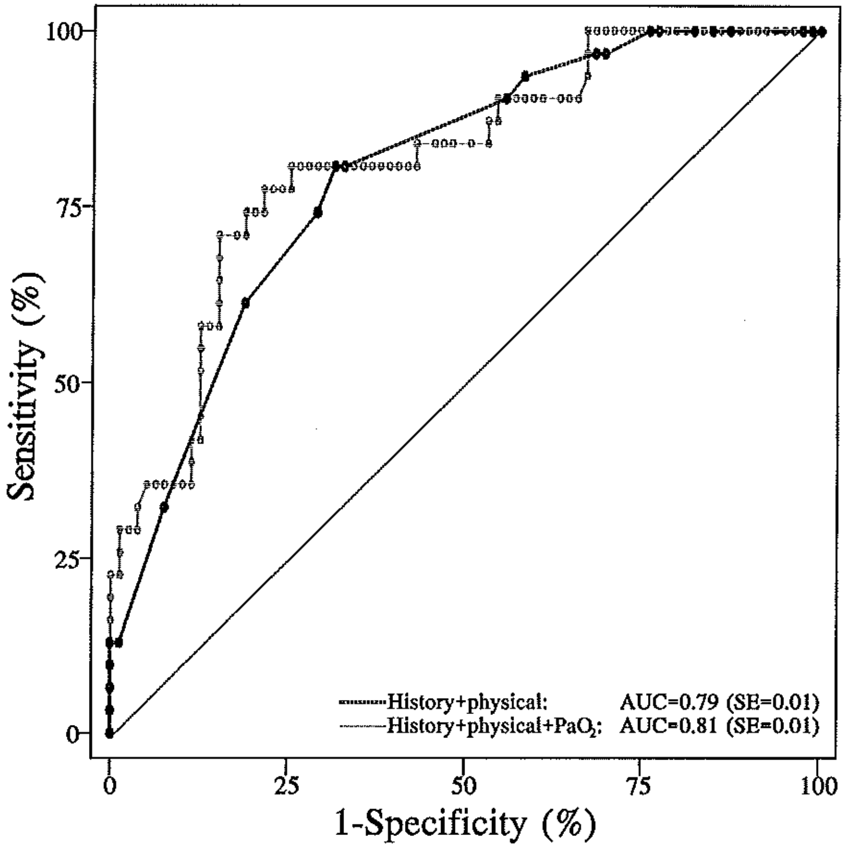
Diagnostic model	Patient history + physical examination		Patient history + physical examination + PaO_2 + X-thorax		Patient history + physical examination + PaO_2 + X-thorax + leg ultrasound	
Determinants	OR	95% CI	OR	95% CI	OR	95% CI
History + physical examination						
Multiple defects on perfusion scan	3.6	1.0- 13.5	5.9	1.4- 26.1	7.3	1.5- 34.5
Cough, new or recently worsened	2.4	0.9- 6.3	2.4	0.8- 6.9	2.1	0.7- 6.6
Previous deep venous thrombosis	13.8	1.4-133.5	23.6	2.2-253.4	13.2	1.1-155.0
Body temperature above 37 °C	3.0	1.1- 8.2	3.4	1.1- 10.3	4.9	1.4- 17.0
Wheezing	0.03	0.00-0.43	0.03	0.00-0.49	0.03	0.00-0.06
Additional diagnostic procedures						
Abnormal chest X-ray	-	-	1.0	1.0- 1.1	1.0	1.0- 1.1
PaO_2 (per mmHg)	-	-	9.5	1.8- 49.7	9.2	1.5- 55.9
Abnormal leg ultrasound	-	-	-	-	9.0	1.4- 55.2

OR, odds ratio; PaO_2, arterial oxygen pressure; CI, confidence interval; °C, degrees Celsius.

model with a combined addition of PaO_2 plus X-ray to the history and physical findings (table 2.2.2) had a ROC area of 0.86. Addition of leg ultrasound to this previous model (table 2.2.2) increased, though not significantly, the ROC area to 0.88. In both diagnostic models, the three tests were associated with the presence of pulmonary embolism although the 95 percent confidence intervals of the odds ratios were wide. Other combinations of the additional procedures, such as leg ultrasonography and chest X-ray only, did not result in a higher diagnostic accuracy.

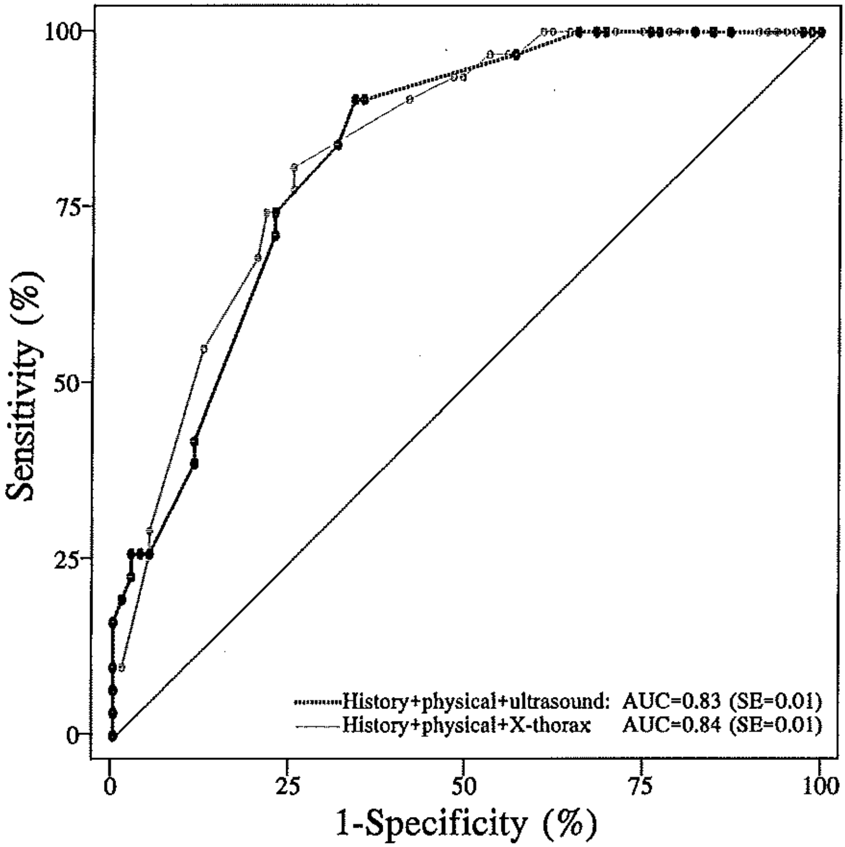
Application of the derived diagnostic models to different patient subsets showed that in the four diseased patients with single defects on the perfusion scan, the patient history and physical examination correctly increased the prior probability of the presence of pulmonary embolism from 0.16 (4/25) to 0.21 (table 2.2.3). Similarly, in diseased patients with multiple defects, the prior increased from 0.32 (27/85) to 0.49. Addition of chest X-ray plus PaO_2 further increased the mean probability of disease. However, this increase was minimal in patients with single perfusion defects.

Figure 2.2.1 The empirical receiver operating characteristic curves of the diagnostic model including patient history and physical examination (multiple defects on perfusion scan, presence of cough, previous deep venous thrombosis, body temperature higher than 37° Celsius and wheezing), and the same model with additional arterial O_2 pressure ($PaO_{2)}$.



AUC = area under the curve; SE = standard error.

Figure 2.2.2 The empirical receiver operating characteristic curves of the diagnostic models including patient history and physical examination (multiple defects on perfusion scan, presence of cough, previous deep venous thrombosis, body temperature higher than 37° Celsius, wheezing) with additional an abnormal X-thorax, and with additional an abnormal leg ultrasound.



AUC = area under the curve; SE = standard error.

In non-diseased patients with single perfusion defects the addition of chest X-ray plus PaO_2 to the patient history and physical examination correctly, though marginally, decreased the prior probability from 0.16 to 0.14 whereas in non-diseased patients with multiple defects they substantially decreased the prior from 0.32 to 0.20. In both

Table 2.2.3 Mean estimated probability of the presence of pulmonary embolism for patients with and without pulmonary embolism per category of number of segment defects on the ventilation scan, estimated by the diagnostic models of Table 2.2.2.

	Patients with pulmonary embolism (n=35)				Patients without pulmonary embolism (n=89)			
Perfusion segment defect	n	Patient history + physical examination	Patient history + physical examination + PaO_2 + X-thorax	Patient history + physical examination + PaO_2 + X-thorax + ultrasound	n	Patient history + physical examination	Patient history + physical examination + PaO_2 + X-thorax	Patient history + physical examination + PaO_2 + X-thorax + ultrasound
Single	4	0.21	0.24	0.27	21	0.15	0.14	0.14
Multiple	27	0.49	0.58	0.60	58	0.24	0.20	0.19

n, number of patients; PaO_2, arterial oxygen pressure.

patient groups, leg ultrasound could neither substantially increase nor decrease the mean probability.

Discussion

Clinical diagnosis is a stepwise process of obtaining information. In subsequent phases diagnostic factors are documented. Commonly, all these factors combined are judged to arrive at a diagnosis which guides treatment. In the present study we therefore constructed a multivariable diagnostic model per work-up phase such that each factor could be evaluated in its clinical perspective and its added value relative to prior information could be judged.

The independent diagnostic determinants of the presence pulmonary embolism were not simply the variables with highest values of test parameters as estimated from univariable analyses. For example, if one of the three tests (PaO_2, X-thorax or leg ultrasound) should be selected based on the univariable approach only, we probably would have chosen leg ultrasound based on the high positive predictive value and likelihood ratio of the positive test result. However, when the three tests were compared in their clinical perspective, i.e. when added to prior information obtained from patient history and physical examination, the three ROC areas did not substantially change which suggests similar diagnostic performance. Moreover, neither in the total patient group nor in the two subgroups leg ultrasound could add diagnostic information to that provided by combined PaO_2, chest X-ray, history, and physical examination. Another example of a misleading clinical potential suggested by single test parameters is that multiple scan defects was an independent predictor for the presence of pulmonary embolism but with a low positive predictive value, positive likelihood ratio and specificity. These parameters were high for pleural rub which, however, appeared to be redundant in the diagnostic process. Similarly, PaO_2 and respiratory frequency had the same ROC area though the latter was no determinant of the presence or absence of pulmonary embolism and did, therefore, not contribute to the diagnosis. These findings as well as the varying added value of PaO_2 plus chest X-ray across patient subgroups, suggest that the clinical relevance of arterial oxygen pressure, chest X-ray and leg ultrasound is determined by other diagnostic. characteristics.

For certain diagnostic variables there were some missing values. This, will not have biased the results as the missing values were equally distributed among the patients with and without pulmonary embolism. However, for PaO_2 there were in total 22 missing values, with twice as many among patients without pulmonary embolism. We, therefore, excluded the PaO_2 test and repeated the multivariable analyses for chest X-ray and leg ultrasound now based on 130 patients. Although the odds ratios of the determinants showed some changes, the inferences from their 95% confidence intervals were the same and so were the ROC areas of the multivariable models.

Although the present analysis is based on limited data, we conclude that patient history, physical examination and other test results are mutually dependent components in clinical diagnosis, i.e. they provide to some extent the same diagnostic information.[15,18-24] For example in our study, previous deep venous thrombosis and advanced age are associated with the presence (and probably also with the severity) of pulmonary embolism. Accordingly, the disease presence (and severity) determines the presence of various symptoms and signs such as cough, multiple perfusion defects, higher arterial oxygen pressure and abnormal X-thorax or leg ultrasound. Therefore, if a disease is commonly diagnosed using more than one test, the clinical relevance of these tests can hardly be judged from univariable measures of association. Although such measures do provide information in a qualitative sense, a 'threshold of diagnostic relevance' cannot be given and will be very arbitrary. If a diagnosis is set by one test, however, the single test parameters can be used to indicate the test's diagnostic potential because mutual dependencies are not applicable. This particularly occurs in screening, in which the early detection of disease irrespective of other clinical characteristics is concerned. We believe that any diagnostic test for a disease outside the realm of screening should be evaluated in its (clinical) context in order to validly assess the added value to information that is recorded regardless.[6,25] This can be realised by multivariable logistic modelling.[26-28] Studies comparing two or more tests that are applied at the same time in the diagnostic work-up in order to replace one test by another for efficiency purposes should be conducted in the clinical context as well. In such studies also the difference in added diagnostic information, or in case of equal effectiveness, the difference in costs or burden to the patient, is clinically relevant. Diagnostic accuracy or test evaluation studies, in which a test is evaluated in isolation should be interpreted with caution as they may give a misleading view of its clinical potential.[6,17,29] It should be realised that in any diagnostic study the mutual dependencies are ignored when considering the results of the reference test (in our example pulmonary angiography).

General standards for (clinical) diagnostic research are needed. Various investigators have proposed a phased approach which resembles the established methodologic standards for evaluation of therapeutic strategies.[1-5,8] Although controversy remains, they propose that if initial diagnostic accuracy studies yield satisfactory test parameters subsequently the test's contribution to the existent diagnostic arsenal in a clinical context should be evaluated, a so-called clinical study.[1-5,7,8] For reasons mentioned before, we believe that the phased approach appropriate for clinical trials on treatment efficacy does not simply apply to diagnostic research. Moreover, diagnostic accuracy studies which mimic a trial Phase One and Two approach (comparing diseased and non-diseased patients as selected on the true disease status) are of very limited value as in such studies the sensitivity, likelihood ratio of a positive test result, ROC area and predictive value tend to be overestimated owing to selection on clinical profile and positive test results.[15,25,30-33] Perhaps only in the case of a completely new test on which no diagnostic information is yet available or the application of an existing test in a completely new clinical context, an initial diagnostic accuracy study (preferably on patients selected on their true disease status) is useful for efficiency reasons. If the test cannot discriminate between the presence and absence of disease the evaluation process could be terminated.[7,8]

Diagnostic research is of great importance and is still in its early phase of development compared to treatment efficacy research. Diagnostic research should provide results that are meaningful to clinical practice. We conclude that parameters such as the sensitivity, specificity, likelihood ratios, ROC area, and predictive value, of single diagnostic tests have limited value. A multivariable logistic regression approach is required to evaluate the independent contribution of each indicator and to construct a diagnostic model.

References

1. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-94.

2. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071-81.

3. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test: lessons from the rise and fall of dexamethasone suppression test. JAMA 1988;259:1699-1702.

4. Köbberling J, Trampisch HJ, Windeler J, eds. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990;28:873-9.

5. Schouw YT van der, Verbeek ALM, Ruijs JHJ. ROC curves for the initial assessment of new diagnostic tests. Fam Pract 1992;9:506-11.

6. Begg CB. Experimental design of medical imaging trials: issues and opinions. Invest Radiol 1989;24:934-6.

7. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental tool in clinical medicine. Clin Chem 1993;39;561-77.

8. Schouw YT van der, Verbeek ALM, Ruijs JHJ. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995;30:334-40.

9. Beek EJR van, Kuyer PMM, Schenk BE, Brandjes DPM, Cate JW ten, Büller HR. A normal perfusion lung scan in patients with clinically suspected pulmonary embolism: frequency and clinical validity. Chest 1995;108:170-3.

10. Beek EJR van, Tiel-van Buul MMC, Büller HR, Royen EA van, Cate JW ten. The value of lung scintigraphy in the diagnosis of pulmonary embolism. Eur J Nucl Med 1993;20:173-81.

11. Michel BC, Seerden RJ, Beek EJR van, Büller HR, Rutten FFH. The cost-effectiveness of diagnostic strategies in patients with suspected pulmonary embolism. Health Economics 1996 (In press).

12. Michel BC, Kuijer PMM, McDonell J, Beek EJR van, Rutten FFH, Büller HR. The derivation of a clinical decision rule in symptomatic pulmonary embolism patients with a non-high probability ventilation-perfusion scan. Submitted.

13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.

15. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978:299;926-30.

16. Begg CB, McNeil BJ. Assessment of radiologic tests; control of bias and other design considerations. Radiology 1988;167:565-9.

17. Swets JA, Getty DJ, Pickettt RM, D'Orsi CJ, Seltzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. Med Decis Making 1991;11:9-18.

18. Sox HC Jr. Probability theory in the use of diagnostic tests. Ann Intern Med 1986;104:60-6.

19. Greenes RA, Begg CB. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. Invest Radiol 1985;20:751-6.

20. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. Am J Cardiol 1986;57:1175-80.

21. Hilden J. Optimistic bias in the assessment of sensitivity and specificity. J Chron Dis 1986;10:853-5.

22. Diamond GA. Clinical epistemology of sensitivity and specificity. J Clin Epid 1992;45:9-13.

23. Pryor DB, Shaw L, McCants CB, Lee KL, Mark DB, Harrell FE Jr, Muhlbaier LH, Califf RM. Value of the history and physical in identifying patients at increased risk for coronary artery disease. Ann Intern Med 1993;118:81-90.

24. Feinstein AR. Clinical judgement revisited: the distraction of quantitative models. Ann Intern Med 1994;120:799-805.

25. Begg CB. Biases in assessment of diagnostic tests. Stat Med 1987;6:411-23.

26. Begg CG, McNeil BJ. Response to "another view of polychotomous analysis". Med Decis Making 1985;5:123-6.

27. Begg CB. Statistical methods in medical diagnoses. CRC Crit Rev Med Inform 1986;1:1-22.

28. Miettinen OS, Caro JJ. Foundations of medical diagnosis: what actually are the parameters involved in Bayes' theorem? Stat Med 1994;13:201-9.

29. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. Invest Radiol 1989;24:934-6.

30. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.

31. Diamond GA. Selection bias and the evaluation of diagnostic tests: a metadissent. J Chron Dis 1986;39:359-60.

32. Knottnerus JA, Leffers JP. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epid 1992;45:1143-54.

33. Schouw YT van der, Dijk R van, Verbeek ALM. 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.

Chapter 2.3

Treatment thresholds in diagnostic test evaluation: an alternative approach to the comparison of areas under the receiver operating characteristic curve

Introduction

To make a diagnosis is not an aim in itself. A diagnosis is relevant in so far as it directs treatment and, therefore, the prognosis of a patient presenting with a particular problem.[1] In clinical diagnosis, the probability of the presence of disease, the availability of additional diagnostic tests and treatment, their risks and benefits, and the severity of the prognosis of the disease when untreated, are imbedded into a complex decision process.[2,3] Most diagnostic studies in applied medical research, however, evaluate the value of diagnostic tests in isolation, rather than in relation to their potential clinical implication for therapeutic decisions. The area under the ROC curve has become the most popular measure of diagnostic accuracy.[2-8] The ROC area, derived either parametrically[9,10] or non-parametrically[11-13], is considered to provide a measure of the overall diagnostic value of the test. It includes the entire range of test results. It is also common practice to compare the overall diagnostic value of two or more (semi)quantitative tests by statistical comparison of their ROC areas.[9,14-17] This, however, may lead to erroneous conclusions in instances where the two ROC curves cross or otherwise have very different shapes.[15] Moreover, a medical decision in practice commonly reflects a dichotomy (to treat or not to treat), or a trichotomy if a subsequent diagnostic test is available (to treat, not to treat or further diagnostic testing). As a consequence, a physician usually operates only at specific parts of the entire range of test results. Several authors have addressed this problem and proposed statistical methods to focus on selected parts of the ROC area, e.g. at fixed sensitivities or specificities.[8,12,17-19] However, little attention has been paid to how to choose these parts.

In this paper we show how evaluation of diagnostic tests irrespective of the clinical application, in particular using overall ROC areas, may give a misleading indication of their clinical relevance. This is done by contrasting this approach with an alternative approach to evaluate diagnostic tests with direct reference to the clinical or therapeutic consequences. Such an approach which applies the basic principles of medical decision making[20-22], has been described by various authors.[2,7,21,23-28]

Example: Patients and ROC analysis

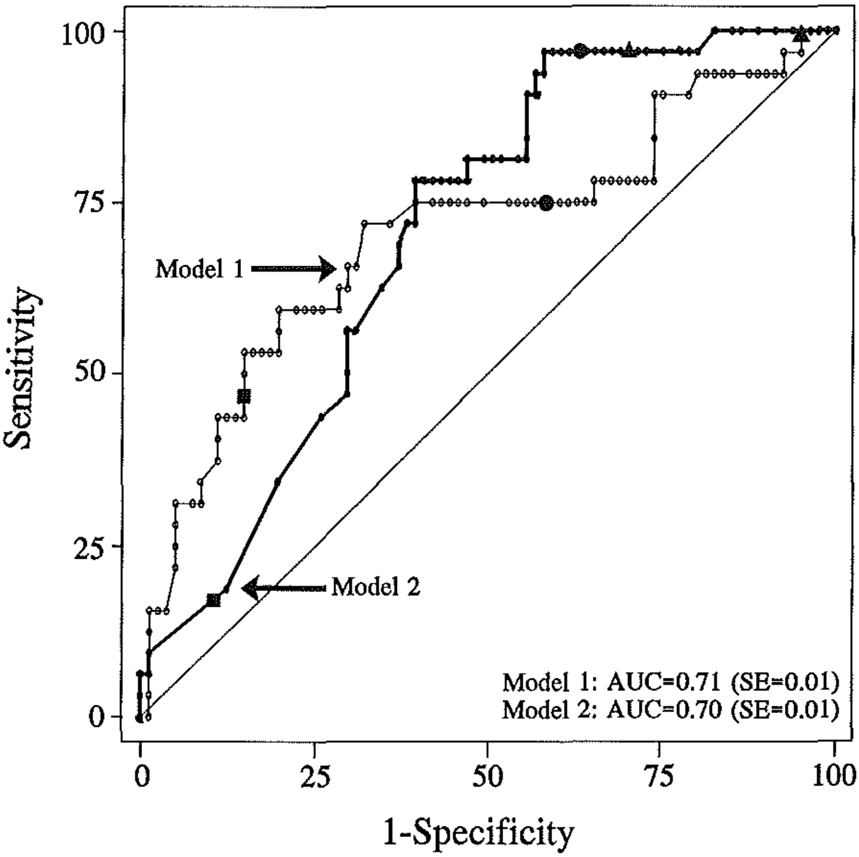
Data were used from consecutive patients with clinically suspected pulmonary embolism who were referred to the Academic Medical Centre and the Slotervaart Hospital in Amsterdam, The Netherlands, as described previously.[29,30] Pulmonary embolism is a lifethreatening disease. Large emboli may cause immediate death. If patients surviving the initial embolism remain untreated, between 18 and 26 percent may die of recurrent embolism.[31,32] The present treatment policy aims to prevent recurrent embolism, and consists of intravenous heparin followed by three to six months of oral anticoagulant therapy, with a risk of serious side effects such as fatal haemorrhage.[33] Therapeutic decisions in the study subjects were based on a negative (do not treat) or high-probability (treat) result of the ventilation-perfusion lung scan. However, 186 of the 452 referred patients had an intermediate non-high probability scan result, and a pulmonary embolism could not be confirmed or excluded. In these patients, patient history, physical examination, routine laboratory tests and pulmonary angiography (reference standard) were obtained prospectively. The pulmonary angiogram was evaluated without knowledge of any other diagnostic information. In this way, the independent contribution of all documented variables to the prediction of presence or absence of pulmonary embolism could be evaluated. In 40 patients pulmonary angiography could not be performed because of medical reasons such as manifest heart failure, severe pulmonary hypertension or poor clinical conditions. In another six patients the angiogram was non-interpretable. Of the remaining 140 patients, 38 patients had an angiographically proven pulmonary embolism ('prevalence' 27 percent).

Using multivariable logistic regression modelling, we constructed various models with different combinations of diagnostic indicators or tests to predict the presence of pulmonary embolism. The diagnostic information content or the discriminative value of the models was compared by their ROC curves. Here, the multivariable logistic regression model is considered as an overall diagnostic test. By entering the observed value of the diagnostic indicators in the logistic function the probability of the presence of a pulmonary embolism was estimated for each patient. Such estimated or post test probability is a probabilistic transformation of the original observed values of the diagnostic indicators in the model. Estimated probabilities can range from 0 to 1 with 0 indicating definitive absence of pulmonary embolism and 1 definitive presence. The area under the ROC curve and its standard error were estimated using a non-parametric approach.[11] The ROC areas of the diagnostic models were compared taking into account the correlation between the models as they were based on the same subjects.[14] The diagnostic models were evaluated on goodness of fit by grouping the patients in ten subgroups according to predicted risk, each subgroup containing an approximately equal number of patients. Per subgroup, the mean of the individual predicted risks was compared with the observed risk using the Hosmer & Lemeshow test statistic.[34] The purpose of this paper is not to report the best discriminating diagnostic model for pulmonary embolism in patients with a non-high probability ventilation-perfusion lung scan, but to demonstrate some pitfalls of conventional diagnostic test (model) comparisons if the clinical context is disregarded. Therefore, we present three illustrative models which would not necessarily be the models of choice for diagnosis of pulmonary embolism, but are selected in order to show the limitations of the use of the ROC area. The models were based on the same 113 patients (27 subjects were excluded due to missing values) of whom 32 had a pulmonary embolism.

Figure 2.3.1 shows the empirical ROC curves of a model that includes the presence of fever (body temperature higher than 37° Celsius), cough and the haemoglobin level in mmol/l (model 1, ROC area = 0.71), and a model including the number of days with symptoms, the number of days of immobilisation, the presence of wheezing and the presence of leg paresis (model 2, ROC area = 0.70). The overall ROC areas, i.e. the area over the entire range of predicted probabilities, were not significantly different (p-value = 0.68) suggesting no difference in diagnostic performance. However, the graph shows crossing ROC curves of the models with different performances in the lower left and upper right part of the graph. Figure 2.3.2 shows the ROC curve of model 1 and a third model including arterial oxygen pressure (mm Hg) and respiratory frequency (model 3, ROC area = 0.61). The ROC areas were significantly different (p-value < 0.001), suggesting that model 1 has better diagnostic properties than model 3. The difference in ROC area was mainly due to the divergence in the middle part of the curves. The Hosmer & Lemeshow test was far from significant (α = 0.05) for all three models which indicates good fit (data not shown).

As usually is the case, only specific parts of the ROC curve are clinically relevant.[2,7,8,12,15,18,19,21,26-28] Inference about diagnostic performance of the three models becomes different when the clinically relevant parts of the ROC curves are considered. This will be evaluated and illustrated below.

Figure 2.3.1 The empirical receiver operating characteristic curve of diagnostic model 1 (body temperature higher than 37° Celsius, presence of cough and haemoglobin level in mmol/l) and diagnostic model 2 (days of symptoms, days of immobilisation, presence of wheezing and presence of leg paresis). The consequences of the three treatment threshold probabilities of pulmonary embolism as estimated from the logistic models are indicated in the figure; 0.33 (■), 0.17 (●) and 0.09 (▲).

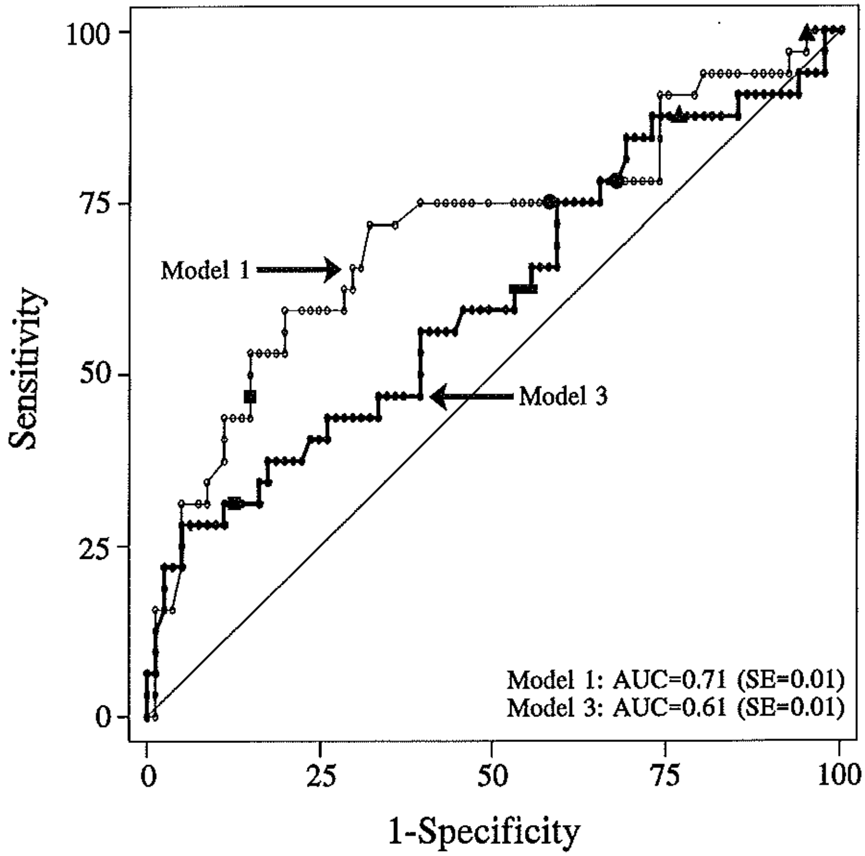


AUC = area under the curve; SE = standard error.

Methodology: Threshold concept in diagnostic test evaluation

To set a diagnosis in clinical practice is to estimate the probability of the presence of disease by the results of diagnostic tests added to the previously obtained information (e.g. patient history and physical examination) in order to decide whether treatment should be initiated or not. Before performing any diagnostic test at all, the prior probability for a patient to have the disease of interest is estimated by the prevalence of the disease in that patient population.

Figure 2.3.2 As figure 2.3.1, except that diagnostic model 2 is replaced with diagnostic model 3 (arterial oxygen pressure in mm Hg and respiratory frequency). The consequences of the three treatment threshold probabilities of pulmonary embolism as estimated from the logistic models are indicated in the figure; 0.33 (■), 0.17 (●) and 0.09 (▲).



AUC, area under the curve; SE, standard error.

By obtaining diagnostic information the ultimate aim is to update this prior probability into a posterior probability of 1 or 0. Because this is usually not feasible, the physician endeavours to increase or decrease the prior probability to justify initiation or withholding of treatment, respectively. To initiate treatment the disease must be present with a sufficiently high probability, i.e. the diagnostic probability must exceed a certain threshold. The acceptable proportion of misclassifications on disease status (false positive and negative diagnosis) and the corresponding risks and benefits of erroneously starting and withholding treatment determine these operating probability thresholds of disease presence.[20-23,27,28] Physicians intuitively apply these thresholds and, thus, implicitly specify the particular range of test results or posterior probabilities that is relevant to clinical decision making. A proper definition and use of these thresholds in diagnostic test evaluation may, therefore, facilitate evaluation of the tests in their clinical perspective.

To reflect clinical practice, we applied the threshold concept of Pauker and Kasirer[20] in analogy to the method described by DeNeef and Kent[25] for evaluation of quantitative diagnostic tests or models. Although this is not necessary, for the sake of simplicity we assume that there are no additional diagnostic tests, and that the objective is to treat or not to treat. The threshold probability above which treatment should be initiated, depends on the disease status and the consequences of treating and not treating. Let $C_{d+,t-}$ denote the consequences (in terms of risks, costs, loss of life time, or, more general, loss of utility) of not treating a patient who has the disease. Let $C_{d+,t+}$, $C_{d-,t+}$ and $C_{d-,t-}$ be defined analogously. If P denotes the patient specific probability of the disease, the expected consequences when the patient is not treated are $P*C_{d+,t-} + (1-P)*C_{d-,t-}$. If the patient is treated the expected consequences are $P*C_{d+,t+} + (1-P)*C_{d-,t+}$. The treatment probability threshold (P_T) is the probability (of the presence of disease) for which the net risks of treatment and no treatment are equal. Pauker and Kasirer showed that this can be written as[20]:

$$P_T = \frac{1}{1 + \dfrac{C_{d+,t-} - C_{d+,t+}}{C_{d-,t+} - C_{d-,t-}}} \qquad (1)$$

A posterior probability of disease presence greater than P_T indicates treatment. P_T reflects the balance between the net "risk" or "cost" of not treating a case with the disease ($C_{d+,t-} - C_{d+,t+}$) and the net "risk" of treating a patient without the disease

($C_{d-,t+}$ - $C_{d-,t-}$).[25] For example, a ratio of net "risks" of 10 means that it is 10 times worse to withhold treatment in a diseased patient than to treat a non-diseased patient. More formally, the ratio of net "risks" is:

$$\frac{C_{d+,t-} - C_{d+,t+}}{C_{d-,t+} - C_{d-,t-}}. \tag{2}$$

Because the terms ($C_{d+,t+}$ - $C_{d+,t-}$) and ($C_{d-,t-}$ - $C_{d-,t+}$) reflect the net "beneficial" and net "deleterious" effects of the treatment, respectively, the ratio in equation 2 has been referred to as the "benefits"/"costs" ratio.[20,25,35,36]

Given a particular treatment threshold P_T and diagnostic test or strategy, the net benefits and risks of subsequent decisions as based on the test when applying that threshold can be estimated. This can be expressed by one parameter, referred to as the expected value[20] or expected utility[25,36] of the diagnostic test. In this paper we will use the term expected risks (ER) in analogy with the above definitions of net risks of treatment and no treatment. The ER of any diagnostic test is equal to:

$$ER = \pi * ER_{d-} + (1 - \pi) * ER_{d-} \tag{3}$$

where π is the prevalence of the disease, and ER_{d+} and ER_{d-} are the expected risks for a patient with and without the disease, respectively. This is equivalent to:

$$\begin{aligned} ER = \pi * [C_{d+,t-} * (1 - \mathit{sensitivity}) + C_{d+,t+} * \mathit{sensitivity}] + \\ (1 - \pi) * [C_{d-,t+} * (1 - \mathit{specificity}) + C_{d-,t-} * \mathit{specificity}] \end{aligned} \tag{4}$$

The sensitivity and specificity of the test can be calculated by estimating the probability of the presence of disease from each observed test result using a logistic model and dichotomising the range of estimated probabilities at the treatment threshold. An estimated probability greater and lower than P_T is defined as a positive and negative test result, respectively.

If there are two diagnostic tests of which the better one has to be selected, the difference in their expected risks, d(ER), at a particular treatment threshold should be compared. This d(ER) can be written as:

$$\begin{aligned} d(ER) = {} & \pi * (C_{d+,t-} - C_{d+,t+}) * (sensitivity1 - sensitivity2) + \\ & (1-\pi) * (C_{d-,t+} - C_{d-,t-}) * (specificity1 - specificity2) \end{aligned} \qquad (5)$$

For comparison of the two diagnostic tests the following index can be used:

$$(sensitivity1 - sensitivity2) + \frac{(1-\pi)}{\pi} * \frac{(C_{d-,t+} - C_{d-,t-})}{(C_{d+,t-} - C_{d+,t+})} * (specificity1 - specificity2) \quad (6)$$

Substituting equation 1 into equation 6 we obtain:

$$(sensitivity1 - sensitivity2) + \frac{(1-\pi)}{\pi} * \frac{P_T}{(1-P_T)} * (specificity1 - specificity2) \qquad (7)$$

Expression 5 and 6 provide an index for comparison of the diagnostic performance of two tests. The index is a kind of weighted comparison (WC) of the sensitivity difference and specificity difference of two tests because it takes into account the relative risks of false positive and negative diagnosis as well as the prevalence of disease. If WC differs from zero the two tests perform differently at the particular treatment threshold or ratio of net "risks". A positive sign is in favour of test 1 whereas a negative sign favours test 2. Since both tests are based on the same patients the standard error of paired proportions can be used to estimate the standard error of WC. Appendix 1 provides a simple method to calculate this standard error.

Example: application of the treatment threshold concept

We applied the above methodology to our example study on diagnosis of pulmonary embolism in order to compare the performance of the three previously constructed diagnostic models. Without explicit definition of the four parameters in equation 2, we arbitrarily defined three ratios of net "risks" at 10, 5 and 2. The corresponding treatment thresholds are 0.09, 0.17 and 0.33, respectively. Under the presumption that the diagnostic models fitted the data correctly, we dichotomised the range of post "test" probabilities of the models at each of these threshold and estimated the corresponding sensitivity, specificity and 95% confidence intervals (table 2.3.1). The points on the ROC curve corresponding to the three posterior probability thresholds are marked in figure 2.3.1 and 2.3.2. Table 2.3.1 and the two figures show that at each threshold the sensitivity and specificity differed to various extents for all three models.

Table 2.3.1 The sensitivity (%), specificity (%), weighted comparison (WC) of the sensitivity and specificity, and their 95% confidence intervals of the three models dichotomised at the treatment threshold of 0.09, 0.17 and 0.33.

Threshold	0.09			0.17			0.33		
	Sensitivity	Specificity	WC*	Sensitivity	Specificity	WC	Sensitivity	Specificity	WC
Model 1	97 (92;100)	6 (1;11)	-	76 (62; 90)	41 (31;51)	-	47 (31;63)	85 (82;94)	-
Model 2	92 (83;100)	28 (20;37)	-0.02 (-0.12;0.08)	92 (83;100)	35 (26;45)	-0.12 (-0.30;0.06)	13 (2;24)	88 (82;94)	0.30 (0.07;0.50)
Model 3	87 (76; 98)	22 (14;30)	0.06 (-0.06;0.18)	79 (66; 92)	30 (21;39)	0.05 (-0.15;0.25)	32 (17;47)	86 (80;92)	0.14 (-0.10;0.39)

* For each treatment threshold, the weighted comparison of model 1 and 2, and model 1 and 3 is presented. Note that a positive WC favors model 1 and a negative WC favors model 2 or 3, respectively.

Model 1 includes: body temperature higher than 37° Celsius, presence of cough and haemoglobin level (mmol/l).

Model 2 includes: days symptoms, days immobilization, presence of wheezing, presence of leg paresis.

Model 3 includes: arterial oxygen pressure, respiratory frequency.

We compared the models at each treatment threshold using the above derived index, WC, and its standard error. The 95% confidence intervals of WC are presented (table 3.2.1). When using the treatment threshold of 0.09, the specificity of model 2 was 22% higher compared to model 1, whereas its sensitivity was 5% lower. Because the balance of "net risks" was 10 to 1, i.e. the loss in sensitivity was much more important than the gain in specificity, model 2 did not perform better than model 1 using the above method (WC = -0.02). A similar result was found in the comparison of model 1 with model 3. At the treatment threshold of 0.17 (ratio of 5 to 1) model 2 had a 16% higher sensitivity and a 6% lower specificity than model 1 with a weighted difference of -0.12 favouring model 2, although the index had a wide confidence interval. However, if the ratio of "net risks" were to be 2 (treatment threshold = 0.33), model 1 performed much better than model 2 (WC = 0.30 in favour of model 1). Both findings are in accordance with the crossing ROC curves (figure 2.3.1) which indicated a similar change in model performance at the two treatment thresholds, whereas the overall ROC areas did not differ. Model 1 and model 3 performed similarly at the treatment threshold of 0.17 (WC = 0.05) whereas they did differ at the threshold of 0.33 (WC = 0.14 in favour of model 1). This also corresponded to the graphical ROC curve presentations (figure 2.3.2).

It should be noted that the key message of table 3.2.1 is the model comparison using the index WC. The 95% confidence intervals are only presented for a proper presentation of the data; with a somewhat larger sample of the patient population the comparison of model 1 and 2 with a WC of -0.12 might be significant in favour of model 2.

Discussion

We have illustrated that statistical inference on the overall ROC area may not reflect the clinical relevance of diagnostic tests or models. A graph may help to reveal this discrepancy. The clinical relevance of a diagnostic test is very much determined by the risks of the misclassifications on disease status and by the prevalence. Application of this knowledge to diagnostic test evaluation may result in different preferences of diagnostic tests compared to when the ROC area method is used. To achieve this, we applied principles suggested in previous studies.[2,7,21,23-28] These studies all proposed in some way to take into account the risks and benefits of subsequent treatment decisions.

We used the treatment threshold concept[20] to define a range of diagnostic probabilities of the presence (and absence) of disease that is relevant to patient management rather than focusing on fixed proportions of the sensitivity or 1-specificity[12] and compared the difference in sensitivity and specificity at the treatment threshold rather than comparing partial ROC areas.[18,19] The plausible range is defined irrespective of the tests and the tests should not necessarily be compared at the same sensitivity, specificity or partial area.[7] We illustrated a simple method to compare, simultaneously, the difference in sensitivity and specificity of two tests (models) given a particular treatment threshold and the 'prevalence' using logistic regression analysis, and to estimate the standard error of this difference. The method may better reflect the difference in test performance as suggested by the graphical ROC presentations, than the statistical comparison of the ROC area.

It should be appreciated that application of the threshold concept to diagnostic test evaluation applies to situations where the ratio of "net risks" or, more general, the benefits/costs ratio can reasonably well be defined. This definition could be based on experience of practising physicians or on general medical knowledge derived from clinical trials and studies on cost-effectiveness.[37-39] In addition to the clinical benefits and risks, economic aspects may also be considered in determining the probability threshold. However, a sensitivity analysis at different thresholds, like we did in our example, remains useful to evaluate whether the choice of the diagnostic test or model changes. Such analyses may also be performed for different prevalences. Methods to evaluate diagnostic tests in situations where these risk and benefits can not be adequately balanced have been described elsewhere.[2,7]

Ideally, diagnostic accuracy is measured independent of the prevalence of the disease.[3,6] However, as the (clinical) performance of a diagnostic test may depend on various factors including the prevalence of the disease[40], we believe it is more appropriate to consider the prevalence in the test evaluation and to pay special attention to the generalisability of study results. It should also be appreciated that different treatment options may have different probability thresholds, depending on their respective benefits/costs ratio. Furthermore, if there are additional diagnostic tests available the category in which the probability is too low to initiate treatment can be further classified in a category of intermediate probability and a category of low probability. In the intermediate category additional tests (in our example a pulmonary angiogram) may be indicated. The lower category, definitive absence of the disease results in discontinuation of diagnostic work-up. In such setting two probability thresholds may be defined.[22] The estimation of the weighted comparison of sensitivity and specificity and its standard error remains essentially unchanged but requires further research.

Evaluation of diagnostic tests without regard of their clinical application may compromise the relevance of the results for medical practice. Recently the importance has been emphasised to conduct diagnostic research within the relevant clinical setting.[41-43] This suggests more pragmatic diagnostic research taking into account their clinical implications. The aim of diagnostic testing is to minimise the uncertainty about the presence or absence of disease in order to reduce the risks of an improper treatment decision. A diagnostic test is clinically relevant if it contributes to this decision[7,44], although there are situations in which the physician applies a diagnostic test only for the sake of knowing what the patient's condition is.[45] In order to agree with practice we believe that benefits and risks of the treatment(s), the untreated prognosis and, if available, of the additional diagnostic tests, should also be considered in diagnostic research or test evaluation. As Pauker and Kassirer stated in 1980[22]: 'The necessity for making such assumptions (about the relative values of benefits and risks of treatment) explicit should be viewed as a strength and not as a weakness of this analytic approach: certainly comparisons of this nature must underlie all clinical decisions, ... '.

References

1. Habbema JDF. Clinical decision theory: the threshold concept. Neth J Med 1995;47:302-7.

2. Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11:95-101.

3. Zweig M, Campbell G. Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561-77

4. Lusted LB. Signal detectability and medical decision making. Science 1971;171:1217-9.

5. Beck JR, Schultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13-20.

6. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285-93.

7. Somoza E, Mossman D. Comparing and optimizing diagnostic tests: an information-theoretical approach. Med Decis Making 1992;12:179-88.

8. Campbell G. General methodology I. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994;13:499-508.

9. Swets JA, Pickett RM. Evaluation of diagnostic systems. Methods from signal detection theory. New York: Academic Press, 1982.

10. Dorfman D, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. J Math Psychol 1969;6:487-96.

11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

12. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984;2:137-50.

13. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234-45.

14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.

15. Centor RM, Schwartz JS. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985;5:149-56.

16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

17. Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, ed. Information processing in medical imaging: proceedings of the eighth conference. The Hague: Martinus Nijhoff, 1984:432-45.

18. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76:585-92.

19. McClish DK. Analyzing à portion of the ROC curve. Med Decis Making 1989;9:190-5.

20. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229-34.

21. Metz CE. Basic Principles of ROC analysis. Semin Nucl Med 1978;8:283-98.

22. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109-17.

23. Greenes RA, Cain KC, Begg CB. Patient-oriented performance measures of diagnostic tests. 1. Tools for prospective evaluation of test order decisions. Med Decis Making 1984;4:7-15.

24. Plasencia CM, Alderman BW, Barón AE, Rolfs RT, Boyko EJ. A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. Med Decis Making 1992;12:204-12.

25. DeNeef P, Kent DL. Using treatment-tradeoff preference to select diagnostic strategies: linking the ROC curve to threshold analysis. Med Decis Making 1993;13:126-32.

26. Halpern EJ, Alpert M, Krieger AM, Metz CE, Maidment AD. Comparisons of ROC curves on the basis of optimal operating points. Acad Radiol 1996;3:245-53.

27. Phelps CE, Mushlin AI. Focusing technology assessment. Med Decis Making 1988;8:279-89.

28. Sainfort F. Evaluation of medical technologies: a generalized ROC analysis. Med Decis Making 1991;11:208-20.

29. Beek EJR van, Kuyer PMM, Schenk BE, Brandjes DPM, Cate JW ten, Büller HR. A normal perfusion lung scan in patients with clinically suspected pulmonary embolism: frequency and clinical validity. Chest 1995;108:170-3.

30. Michel BC, Seerden RJ, Beek EJR van, Büller HR, Rutten FFH. The cost-effectiveness of diagnostic strategies in patients with suspected pulmonary embolism. Health Economics 1996 (In press).

31. Barrit DW, Jordan SC. Anticoagulant drugs in the treatment of pulmonary embolism, a controlled trial. Lancet 1960;1309-12.

32. Coon WW, Willis PW, Keller JB. Venous thrombosis and other venous diseases in the Tecumseh community health study. Circulation 1973;48:839-46.

33. Levine MN, Raskob G, Hirsh J. Hemorrhagic complications of long-term anticoagulant therapy. Chest 1989;95 suppl:26S-36S.

34. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc, 1989:140-5.

35. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: W.B. Saunders, 1980:121-6.

36. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworths, 1988:134-45.

37. Bernstein SJ, Hilborne LH, Leape LL, et al. The appropriateness of use of coronary angiography in New York state. JAMA 1993;269:766-9.

38. Kassirer JP. The quality of care and the quality of measuring it. N Engl J Med 1993;329:1263-5.

39. Tanenbaum SJ. What physicians know. N Engl J Med 1993;329:1268-71.

40. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.

41. Evidence-Based Medicine Working Group. Evidence-based medicine. JAMA 1992;268:2420-5.

42. Epstein AM. Sounding Board. The outcomes movement-will it get us where we want to go? N Engl J Med 1990;323:266-70.

43. Grobbee DE, Miettinen OS. Clinical epidemiology: introduction to the discipline. Neth J Med 1995;47:2-5.

44. Ash DA, Patton JP, Hershey JC. Knowing for the sake of knowing: the value of prognostic information. Med Decis Making 1990;10:47-57.

45. Berwick DM, Weinstein MC. What do patients value? Willingness to pay for ultrasound in normal pregnancy. Med Care 1985;23:881-93.

Appendix

To evaluate whether two diagnostic tests perform differently at a particular treatment threshold, a confidence interval around the weighted comparison (WC) of their sensitivities and specificities can be estimated. Formally, WC is:

$$(sensitivity1 - sensitivity2) + \frac{(1-\pi)}{\pi} * \frac{P_T}{(1-P_T)} * (specificity1 - specificity2). \quad \text{(A.1)}$$

The odds for the prevalence π and for the treatment threshold P_T are both independent of the test and will be regarded as constants, though it should be noted that the odds for the prevalence also has a confidence interval. Because both sensitivities and specificities are based on the same diseased and non-diseased patients, respectively, the standard error (SE) of WC equals

$$\sqrt{}\,[\ SE(sensitivity1 - sensitivity2)^2 + C^2 * SE(specificity1 - specificity2)^2\ ]. \quad \text{(A.2)}$$

This standard error can easily be computed:

Fit a suitable model (in the present application, a logistic model) for each test. Let P be the estimated probability of the presence of disease by test 1 (P1) and test 2 (P2). Subsequently, $P \geq P_T$ is defined as a positive test result (+) and $P < P_T$ as a negative test result (-). Compute difference D as follows:

D = 1 if P1 = + and P2 = -
D = -1 if P1 = - and P2 = +
D = 0 if P1 = P2.

Given the diseased patients, the mean and the standard error of D represent the difference and the standard error of the difference in sensitivity of the two tests. Given the non-diseased patients, the mean and the standard error of D represent the difference and the standard error of the difference in specificity of the two tests. Substituting both standard errors into equation A.2 provides the standard error of the weighted comparison WC (equation A.1).

CHAPTER 3

APLLICATION

Chapter 3.1

Evaluation of the independent diagnostic determinants of pulmonary embolism among patient history, physical examination, blood gas analysis, chest X-ray and perfusion lung scintigraphy

Introduction

Diagnosis in patients with clinically suspected embolism is a difficult task. Leaving a pulmonary embolism untreated may be fatal in approximately 20% of the patients[1,2], whereas the current treatment strategy (intravenous heparin followed by three to six months of anticoagulantia) may cause fatal haemorrhage.[3] Therefore, an immediate and proper diagnosis in these patients is required. The diagnostic work-up in patients with clinically suspected pulmonary embolism consists successively patient history, physical examination, blood gas analysis and chest X-ray. When the results of this work-up remains suggestive for pulmonary embolism perfusion lung scintigraphy and, in the case of an abnormal perfusion scan, ventilation lung scanning are performed.[4-7] However, even though this diagnostic sequence is common practice in many hospitals the added or independent value of the separate tests to previously obtained diagnostic information has not been investigated.

This study evaluates the independent diagnostic value of patient history, physical examination, arterial blood gas values and chest X-ray in the diagnosis of pulmonary embolism, using multivariable diagnostic (logistic regression) models. Subsequently, the added value of perfusion lung scan results is evaluated.

Patients and Methods

Patients

The study population comprised 451 consecutive patients with clinically suspected pulmonary embolism who were referred to the Academic Medical Centre and the Slotervaart Hospital in Amsterdam, The Netherlands, between April 1991 and October 1993.[5,8,9] To set a diagnosis in these patients, the history (e.g. age, dyspnoea, previous deep venous thrombosis (DVT) and recent surgery), physical examination (e.g. fever, pleural rub and respiratory frequency), blood gas values (partial pressure of oxygen and carbon dioxide in arterial blood, PaO_2 and $PaCO_2$), chest radiograph, perfusion and ventilation scan results were obtained. Generally, a normal perfusion scan is considered as evidence for the absence of pulmonary embolism (no treatment indication)[4,10], and a segmental or larger perfusion defect in combination with a normal ventilation scan, i.e. a high probability ventilation-perfusion scan result for its presence.[4,11-13] These angiographically controlled studies have shown that more than 90% of the high-probability patients has pulmonary embolism. Therefore, treatment is commonly initiated in these patients. In the current study, patients with an inconclusive, i.e. a non-high probability ventilation-perfusion scan underwent pulmonary angiography to determine the presence of pulmonary embolism. Accordingly, in this study, pulmonary embolism was considered present by a an abnormal angiogram or a high probability ventilation-perfusion scan and absent by a normal angiogram or normal ventilation-perfusion scan. The two lung scans and the angiograms were independently evaluated without knowledge of any other diagnostic information. The chest X-ray was considered abnormal if it showed an elevated hemidiaphragm, a small pleural effusion, atelectasis or parenchymal abnormalities (consolidation). After the recording of the patient history, physical examination, routine laboratory and chest X-ray, but before lung scanning, the same physician was asked to score the patient's probability of having pulmonary embolism as low (< 10%), medium low (10% to 50%), medium high (50% to 90%) or high (> 90%).

Analysis

Data analysis was performed with standard software packages (SAS Institute Inc., Cary release 6-10). Differences and 95% confidence intervals (CI) in frequencies or mean values of all diagnostic variables among patients with and without pulmonary embolism were calculated. In accordance with the sequence of diagnostic work-up in clinical practice, we initially included all potential and clinically relevant diagnostic determinants obtained from the patient history in a multivariable logistic regression model. The diagnostic information content or the discriminative value of various reduced models was compared with the overall model using the area under the Receiver Operating Characteristic (ROC) curve. The ROC area and its standard error were estimated using the non-parametric approach.[14] In the model comparisons, the correlation between models was taken into account because they were based on the same subjects.[15] This model reduction was done to obtain the most efficient diagnostic model, i.e. the model with a minimum of determinants that did not have a significantly lower ROC area than the overall model. The same approach was applied to all physical examination findings after they were added to the most efficient "patient history model". This allowed to evaluate whether data from physical examination had independent diagnostic value, i.e. added to the patient history which is always obtained first, and if so which physical findings determine this incremental diagnostic value. Similarly, findings of the blood gas analysis, chest X-ray and lung perfusion scan were consecutively added to each previous model, to evaluate their incremental value in the diagnosis of pulmonary embolism. All continuous variables were included into the models uncategorised if a linear relation was plausible.

The reliability of the diagnostic models was evaluated by grouping the patients in ten subgroups according to predicted risk, each subgroup containing an approximately equal number of patients. Per subgroup, the mean of the individual predicted risks was compared with the observed risk using the Hosmer & Lemeshow test statistic.[16] Several authors have suggested to evaluate (differences in) diagnostic test performance across clinically different patient subsets.[17-19] Therefore, we applied the models to the four patient subgroups with increasing clinical probability of pulmonary embolism as estimated by the patient's physician before lung scanning. Per subgroup, we compared the mean predicted probabilities of the diagnostic models. This provided both a kind of validation study of the different models in different patient subsets and a possibility to compare the physician's estimated probability with the observed probability and predicted probability of pulmonary embolism.

Results

Of the 451 patients, 126 (28%) had a normal, 132 (29%) a high-probability, 186 (41%) a non-high probability ventilation-perfusion scan and in 7 (2%) patients these tests were not performed due to the finding of an abnormal test for deep vein-thrombosis. In 40 of the 186 patients with a non-high probability scan, pulmonary angiography could not be performed because of medical reasons such as manifest heart failure, severe pulmonary hypertension or poor clinical conditions. In another six patients the angiogram was non-interpretable. Of the remaining 140 patients, 38 patients (27%) had an angiographically proven pulmonary embolism. In total, 398 patients were suitable for analyses of which 170 were considered to have pulmonary embolism (overall prevalence: 43%).

Of patient history and physical examination, age, days of immobilisation, presence of malignancy, surgery within past 3 months, circulatory collapse, dyspnoea, previous deep venous thrombosis, leg paresis, signs of deep venous thrombosis, pleural rub,

Table 3.1.1 Association between various patient characteristics and the presence of pulmonary embolism among 398 patients suspected of pulmonary embolism.

Diagnostic variables	Pulmonary embolism present (N=170)	Pulmonary embolism absent (N=228)	Difference (95% CI)
Patient history			
Age (years)	59.4*	52.9*	6.5 (3.0; 9.9)
Sex (% male)	45	42	3 (- 7;13)
Days of immobilization	0 (0-7)†	0 (0-3)†	‡
Malignancy (%)	30	19	11 (2;19)
Surgery within past 3 months (%)	28	16	12 (4;20)
Family history of thrombosis (%)	9	11	-1 (- 7; 5)
Collapse (%)	13	4	9 (3;15)
Dyspnoea (%)	18	30	-12 (-20;-4)
Previous DVT (%)	11	5	6 (1;11)
Previous embolism (%)	8	7	1 (- 5; 6)
Palpitations (%)	18	16	2 (- 5; 9)
Physical examination			
Body mass index (kg/m^2)	24.6*	24.6*	0 (- 1; 1)
Leg paresis (%)	8	4	4 (- 1; 9)
Signs of DVT (%)	12	7	5 (- 1;11)
Pleural rub (%)	20	13	7 (1;13)
Body temperature > 37 °C (%)	46	39	7 (- 3;17)
Respiratory frequency (breaths/min)	21*	19*	2 (0; 4)
Heart rate (beats/min)	95*	91*	4 (- 0; 8)
Additional tests			
Arterial O_2 pressure (mm Hg)	73.8*	74.6*	-0.8 (-5.4;3.8)
Arterial CO_2 pressure (mm Hg)	35.5*	36.0*	-0.5 (-2.2;1.2)
Abnormal chest X-ray (%)	51	32	19 (9; 29)
Perfusion scan			
normal (%)	1	54	-53 (-59;-47)
subsegmental defect (%)	10	20	-10 (-17;- 3)
segmental or larger defect (%)	89	26	63 (56; 70)

n, number of patients; DVT, deep venous thrombosis; min, minute.

* mean

† median with 25th and 75th percentiles between parentheses

‡ p-value < 0.001 according to the Mann Whitney rank sum test.

body temperature above 37°C, respiratory frequency, and heart rate showed substantial differences for patients with and without pulmonary embolism (table 3.1.1). PaO_2 en $PaCO_2$ were equally distributed in both patient groups. 156 patients (39%) had an abnormal chest X-ray of which 86 (55%), 70 (45%), 43 (28%) and 21 (13%) had a small pleural effusion, consolidation, an elevated hemidiaphragm or atelectasis, respectively. The percentage of abnormal X-rays was 19% (95% CI: 9%-29%) higher in patients with pulmonary embolism as compared to those without. 129 (33%) patients had a normal perfusion lung scan, 61 (15%) a subsegmental defect and 208 (52%) a segmental or larger defect. Subsegmental defects were much more prevalent in patients without pulmonary embolism. Segmental or larger defects were more frequent (63%, 95% CI: 56%-70%) in patients with pulmonary embolism.

The multivariable analysis was based on 360 patients. 38 patients were excluded due to missing values. The overall diagnostic model including all relevant variables of patient history had a ROC area of 0.69. A reduced model including age, surgery within past 3 months, previous deep venous thrombosis, dyspnoea, collapse and malignancy (table 3.1.2) had a ROC area of 0.68 (figure 3.1.1). Further exclusions, as well as other combinations of patient history factors, significantly decreased the ROC area. Therefore, malignancy remained in the model although the 95% CI of the odds ratio just included 1.0. Addition of all relevant physical examination findings to the previous, i.e. the most efficient patient history model significantly increased the ROC area from 0.68 to 0.72. However, excluding all physical findings except pleural rub, signs of deep venous thrombosis and respiratory frequency from this model yielded also a ROC area of 0.72 (figure 3.1.1). These three factors were the only physical findings that predicted the presence of disease independently from patient history, although pleural rub was borderline significant (p-value = 0.09). In this model all history findings remained independent predictors as well (table 3.1.2). Addition of the two blood gas parameters to the previous reduced model including history plus physical findings did not increase the ROC area (figure 3.1.1) and both had no independent association with pulmonary embolism. The odds ratios (OR) of PaO_2 and $PaCO_2$ were 1.0 (95% CI: 0.99-1.01) and 1.0 (95% CI: 0.96-1.03), respectively. Although an abnormal chest X-ray, when added to the patient history and physical examination, showed an independent relation with pulmonary embolism (OR = 2.3, table 3.1.2), the increase in ROC area was low (from 0.72 to 0.74, figure 3.1.2). Addition of the perfusion scan result (included as a dichotomous variable with no or

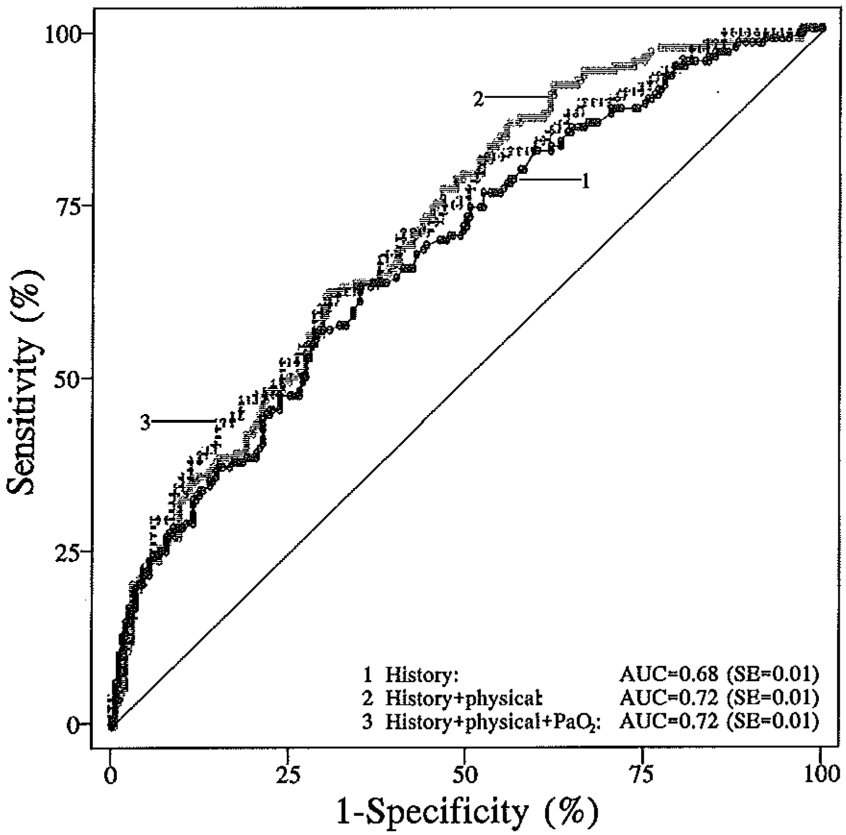
Table 3.1.2 Results of the multivariable logistic regression analysis for four diagnostic models including patient history, additional physical examination, additional chest X-ray and additional perfusion scan results, to assess the presence of pulmonary embolism in 360 patients suspected of pulmonary embolism.

Diagnostic model	History[1]	History + Physical[2]	History + Physical + Chest X-ray[3]	History + Physical + Chest X-ray + Perfusion scan[4]
Determinants	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Patient history				
Age (10 years)	1.1 (1.0-1.3)	1.1 (1.0- 1.3)	1.1 (1.0- 1.3)	1.0 (0.9- 1.2)
Surgery within past 3 months	2.0 (1.2-3.5)	1.9 (1.1- 3.4)	1.8 (1.0- 3.1)	1.4 (0.7- 2.8)
Previous DVT	2.8 (1.1-6.7)	2.5 (1.0- 6.2)	2.6 (1.0- 6.6)	2.6 (0.9- 7.4)
Dyspnoea	0.5 (0.3-0.9)	0.6 (0.3- 1.0)	0.6 (0.3- 1.0)	0.9 (0.4- 1.8)
Collapse	3.7 (1.5-9.4)	3.9 (1.6-10.0)	4.9 (1.9-12.8)	3.7 (1.2-11.4)
Malignancy	1.4 (0.9-2.4)	1.4 (0.8- 2.4)	1.3 (0.8- 2.2)	1.0 (0.5- 1.9)
Physical examination				
Pleural rub		1.7 (0.9- 3.2)	1.3 (0.7- 2.5)	1.0 (0.5- 2.2)
Signs of DVT	-	2.1 (1.0- 4.5)	2.2 (1.0- 4.8)	1.8 (0.7- 4.4)
Respiratory frequency (10 breaths/min)	-	1.3 (1.0- 1.7)	1.2 (0.9- 1.6)	1.3 (0.9- 1.7)
Additional tests				
Abnormal chest X-ray	-	-	2.3 (1.4- 3.8)	1.5 (0.8- 2.6)
Segmental or larger	-	-	-	18.2 (9.5-34.9)

OR, odds ratio; CI, confidence interval; DVT, deep venous thrombosis; min, minute.

[1] Baseline odds = 0.5; [2] Baseline odds = 0.2; [3] Baseline odds = 0.2; [4] Baseline odds = 0.05.

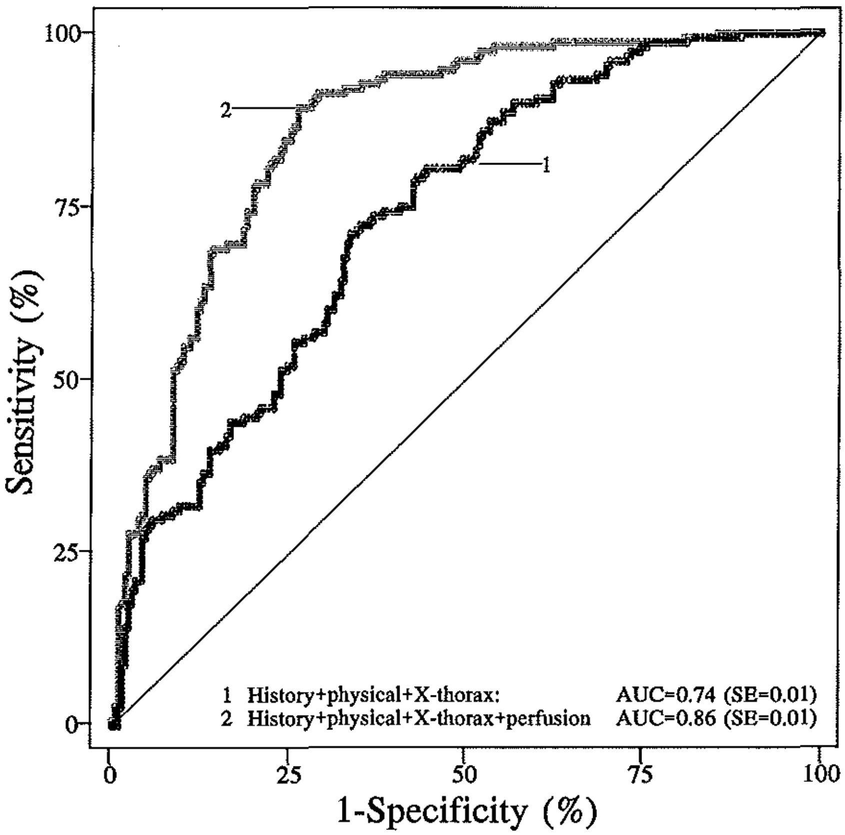
Figure 3.1.1 The empirical receiver operating characteristic curves of the diagnostic model including patient history, the model including patient history and physical examination, and the model including patient history, physical examination and arterial O_2 pressure (PaO_2).



AUC = area under the curve; SE = standard error.

subsegmental perfusion defect as the reference category) to the most efficient history-physical-chest X-ray model increased the ROC area from 0.74 to 0.86 (figure 3.1.2). In this model, the perfusion scan was a very strong independent determinant (OR = 18.2) of the presence of pulmonary embolism. The odds ratio of chest X-ray and various history and physical findings substantially decreased (table 3.1.2). After exclusion of chest X-ray findings, the ROC area remained 0.86. The Hosmer & Lemeshow test was far from significant for each model which indicates well fitted models (data not shown).

Figure 3.1.2 The empirical receiver operating characteristic curves of diagnostic model including patient history, physical examination and X-thorax, and the diagnostic model including patient history, physical examination, X-thorax and perfusion lung scan.



AUC = area under the curve; SE = standard error.

Table 3.1.3 Mean estimated probability of the presence of pulmonary embolism by the three derived diagnostic models including patient history and physical examination (table 3.1.2), additional chest X-ray (table 3), and additional perfusion scan (table 3) for all patients with and without emboly and for four patient risk groups as defined by the patient's physician prior to the ventilation-perfusion lung scanning.

	Patients with pulmonary embolism				Patients without pulmonary embolism			
Risk	n	PH + PE	PH + PE + chest X-ray	PH + PE + perfusion scan	n	PH + PE	PH + PE + chest X-ray	PH + PE + perfusion scan
Low	3	0.50	0.50	0.72	24	0.24	0.24	0.20
Medium low	50	0.47	0.51	0.61	120	0.34	0.33	0.21
Medium high	72	0.49	0.50	0.65	60	0.43	0.42	0.34
High	23	0.54	0.59	0.72	8	0.39	0.35	0.23
Total	148	0.49	0.51	0.65	212	0.36	0.35	0.24

n = number of patients; PH = patient history; PE = physical examination.

Table 3.1.3. shows the results of application of three diagnostic models to the total patient group on which the models were based, and to different subgroups. The prevalence, or prior probability, of pulmonary embolism was 41% (148/360). On average this prior could correctly be increased and decreased by patient history and physical examination, in patients with and without pulmonary embolism, respectively. In accordance with previous multivariable analyses, addition of chest X-ray did not increase or decrease the mean probability of disease, in contrast to addition of perfusion scan results. Subgroup analysis yielded similar results. In patients with a low (< 10%), medium low (10%-50%), medium high (51%-90%) and high probability (> 90%) as estimated by their physician before lung scanning, the observed prevalences were 11% (3/27), 29% (50/120), 55% (72/132) and 74% (23/31), respectively. In the diseased subjects of each subgroup with exception of the high risk group, the respective priors were correctly increased by addition of perfusion scan results. In the non-diseased subjects the reverse was found: only in the high, medium high and medium low risk group the perfusion scan findings correctly decreased the prior probabilities.

Discussion

This study examined the value of patient history and the added value of physical examination, arterial blood gas analysis, chest radiography and perfusion scintigraphy, in the diagnosis of pulmonary embolism, according to the usual order in which these data become available in clinical practice. This was done by systematically constructing and extending multivariable diagnostic models. We found that the independent predictors obtained from patient history (age, recent surgery, previous deep venous thrombosis, dyspnoea, collapse) and physical examination (pleural rub, signs of deep venous thrombosis, respiratory frequency) contribute to the confirmation or exclusion of the presence of pulmonary embolism. Given this information from patient history and physical examination, blood gas measurement has no added diagnostic value and the added value of chest radiography is limited.

Our results suggest that in the assessment of the presence or absence of pulmonary embolism, patient history and physical examination comprised all diagnostic information to be obtained from the arterial blood gas and to a large extent the information from the chest radiograph. Moreover, addition of the perfusion scan results makes the chest radiograph redundant to determine the presence or absence of pulmonary embolism. As both blood gas and chest radiography measurements are more or less burdening for the patient and expensive, refraining from these examinations may increase the efficiency in diagnosis of pulmonary embolism. However, chest radiography may be of help in the assessment of other lung diseases when pulmonary embolism is excluded. Furthermore, it should be recognised that after the addition of perfusion scan results, history and physical findings such as age, dyspnoea, malignancy, pleural rub and respiratory frequency also contribute much less to the diagnosis of pulmonary embolism. Patient history and physical examination are, however, always obtained, these should obviously not be disregarded in the diagnostic work-up.

The results of this study regarding the predictors of pulmonary embolism agree with results from earlier studies.[20-24] However, in previous studies the value of the above diagnostic variables or tests was examined in isolation or in a univariable sense, without reference to diagnostic information already available. In agreement with previous reports, the present study shows a univariable association of immobilisation, tachycardia and radiographic abnormalities with the presence of pulmonary embolism whereas an association independent from other history and physical findings could not be found. This is probably due to a mutual dependency with these other stronger predictors. For example, the diagnostic information of immobilisation may largely overlap with the information of recent surgery. With respect to blood gas values the results of this study are different from some other studies[25,26], which is most likely explained by the difference in study design and because we used an independent analysis of the blood gas values.

The subgroup analyses in the present study suggested that the added diagnostic value of the perfusion lung scan for confirming pulmonary embolism was most profound in patients with low, medium low or medium high risk and its added value for excluding the disease in high, medium high and medium low risk patients. Although this may clinically be expected, i.e. the potential of the perfusion scan to increase or decrease the prior will be less if the prior is already high or low, it should be realised that the groups and therefore the corresponding priors were based on the risk estimated the physician before lung scanning. In the low and medium low risk group, the physician's estimated risk agreed with the true prevalence of pulmonary embolism but in the medium high and high risk group the physician tended to overestimate the disease probability, particularly in the high risk group. Nevertheless, the overall added value of the perfusion scan as found in the present study is in accordance with earlier findings.[4,5,7,12,13] The large contribution of the perfusion scan in our study can be explained by the fact that a normal perfusion scan and a high probability ventilation-perfusion scan result define the final diagnosis of pulmonary embolism. This definition of the presence and absence of pulmonary embolism, however, accords to prevailing clinical practice. Currently, patients with a normal perfusion scan are considered to be free of pulmonary embolism (not treated) and a patient with high-probability ventilation-perfusion scan is directly treated with anticoagulants without further diagnostic evaluation. Our pragmatic disease definition may also partly explain the high prevalence in this study (43%) as compared to previous studies with prevalences of approximately 30%.[4,12] The present study included relatively more high-probability patients which also may account for a higher prevalence. We believe that the high prevalence has not influenced the observations and conclusions. Moreover, variables which are theoretically associated with the presence of pulmonary embolism have a bigger chance to be detected if the prevalence is high. Therefore, the limited diagnostic value of arterial blood gas values and chest radiography in establishing or excluding pulmonary embolism when prior information is available remains.

The model including history, physical examination and chest X-ray could not distinguish between the different risk groups as well as the physician. The physician may have used more information for the risk estimation than was included in the model. Future studies may evaluate the effects of a more prudent interpretation of patient history and physical examination on the physician's assessment of the (prior) probability of presence of pulmonary embolism.

In conclusion, we have shown that efficiency in establishing the diagnosis of patients with suspected pulmonary embolism may increase if the sequence of the diagnostic work-up is evaluated. Doing so, blood gas analysis and chest radiography may become redundant to the patient history and physical examination in a strategy aimed at diagnosing or excluding pulmonary embolism.

References

1. Barrit DW, Jordan SC. Anticoagulant drugs in the treatment of pulmonary embolism, a controlled trial. Lancet 1960;1309-12.

2. Coon WW, Willis PW, Keller JB. Venous thrombosis and other venous diseases in the Tecumseh community health study. Circulation 1973; 48:839-46.

3. Levine MN, Raskob G, Hirsh J. Haemorrhagic complications of long-term anticoagulant therapy. Chest 1989;95 suppl:26S-36S.

4. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). JAMA 1990;263:2753-9.

5. Beek EJR van, Tiel-van Buul MMC, Büller HR, Royen EA van, Cate JW ten. The value of lung scintigraphy in the diagnosis of pulmonary embolism. Eur J Nucl Med 1993;20:173-81.

6. Stein PD, Hull RD, Saltzman HA, Pineo G. Strategy for diagnosis of patients with suspected acute pulmonary embolism. Chest 1993;103:1553-9.

7. Gottschalk A, Juni JE, Sostman HD, et al. Ventilation-perfusion scintigraphy in the PIOPED study. Part II. Evaluation of the scintigraphic criteria and interpretations. J Nucl Med 1993;34:1119-26.

8. Beek EJR van, Kuyer PMM, Schenk BE, Brandjes DPM, Cate JW ten, Büller HR. A normal perfusion lung scan in patients with clinically suspected pulmonary embolism: frequency and clinical validity. Chest 1995;108:170-3.

9. Michel BC, Seerden RJ, Beek EJR van, Büller HR, Rutten FFH. The cost-effectiveness of diagnostic strategies in patients with suspected pulmonary embolism. Health Economics 1996 (In press).

10. Hull RD, Raskob GE, Coates G. Panju AA. Clinical validity of a normal perfusion lung scan in patients with suspected pulmonary embolism. Chest 1990;97:23-6.

11. Biello DR, Mattar AG, McKnight RC, Siegel BA. Ventilation-perfusion studies in suspected pulmonary embolism. AJR 1979;133:1033-7.

12. Hull RD, Hirsh J, Carter CJ, et al. Diagnostic value of ventilation-perfusion scan in patients with suspected pulmonary embolism. Chest 1985:88;819-28.

13. Hull RD, Raskob GE, Coates G, et al. A new non-invasive management strategy for patients with suspected pulmonary embolism. Arch Intern Med 1989;149:2549-55.

14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

15. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-843.

16. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc; 1989:140-5.

17. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978:299;926-30.

18. Begg CB, McNeil BJ. Assessment of radiologic tests; control of bias and other design considerations. Radiology 1988;167:565-9.

19. Swets JA, Getty DJ, Picket RM, D'Orsi CJ, Seltzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. Med Decis Making 1991;11:9-18.

20. Stein PD, Terrin ML, Hales CA, et al. Clinical, Laboratory, Roentgenographic and electrocardiographic findings in patients with acute pulmonary embolism and no pre-existing cardiac or pulmonary disease. Chest 1991;100:598-603.

21. Stein PD, Saltzman HA, Weg JG. Clinical characteristics of patients with acute pulmonary embolism. Am J Cardiol 1991;68:1723-4.

22. Stein PD, Gottschalk A, Saltzman HA, Terrin ML. Diagnosis of acute pulmonary embolism in the elderly. Am J Coll Cardiol 1991;18:1452-7.

23. Stein PD, Willis PW III, DeMets DL, Greenspan RH. Palin chest roentgenogram in patients with acute pulmonary embolism and no preexisting cardiac or pulmonary disease. Am J Noninvas Cardiol 1987;1:171-6.

24. Driel AD van, Ullmann EF, Bosch FH. Arteriële-bloedgasanalyses bij long-embolie; soms een luchtspiegeling (in Dutch). NTVG 1992;136:305-8.

25. Cvitanic O, Marino PL. Improved use of arterial blood gas analysis in suspected pulmonary embolism. Chest 1989;95:48-51.

26. Szucs MM, Brooks HL, Grossman W, et al. Diagnostic sensitivity of laboratorium findings in acute pulmonary embolism. Ann Intern Med 1971;74:161-6.

Chapter 3.2

Continuous ST-segment monitoring to predict infarct size and left ventricular function in the GUSTO-I trial

The present study evaluates whether continuous ST-monitoring characteristics are associated with the enzymatic infarct size and left ventricular ejection fraction (LVEF) in patients with acute myocardial infarction. Both measures of disease severity (infarct size and LVEF) were studied as quantitative rather than as dichotomous parameters. Accordingly, we used linear regression analysis to quantify the above associations. This analysis should be regarded as an initial approach to evaluate the value of the continuous ECG-monitoring test for prediction of the myocardial infarct size and left ventricular function. However, to evaluate its true value from a diagnostic (and prognostic) perspective in order to guide subsequent treatment decisions, a certain threshold on the disease parameters must be defined. As this is still a rather ambiguous issue in the medical literature, this was not attempted in the present study. Analysis of the data using diagnostic (logistic) modelling has therefore not yet been attempted. Further research and determination of clinically relevant cut-off levels is required to direct such an evaluation.

Introduction

Infarct size is a major determinant of the prognosis of patients with an acute myocardial infarction.[1] Early reperfusion (within a few hours) and sustained patency of the occluded artery by thrombolytic therapy limits infarct size and, thereby, preserves left ventricular function and improves survival.[2-7] Infarct size and left ventricular function are associated with the extend and duration of the myocardial ischaemia. Continuous monitoring of the ST-segment is a readily available, non-invasive method for assessment of myocardial ischaemia and the occurrence of early reperfusion and reocclusion(s).[8-13] Hence, this technique may well provide diagnostic information of both infarct size and residual left ventricular function at the early stages of myocardial infarction.

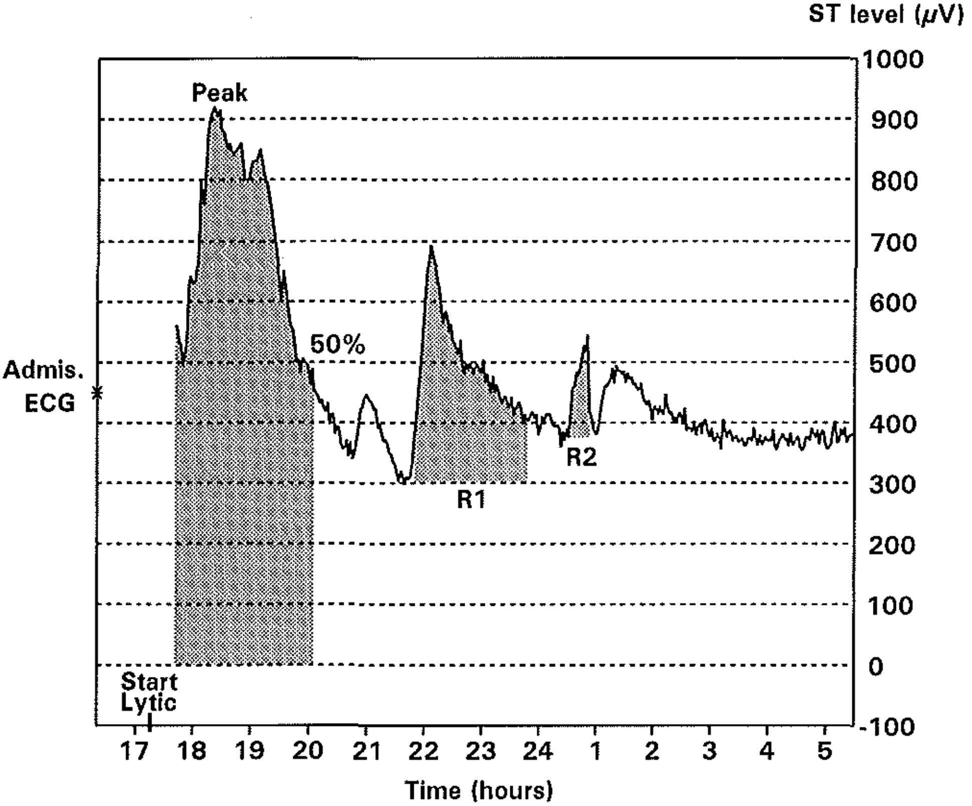
The ECG monitoring substudy of the GUSTO-I trial[13,14] offers an unique opportunity to verify these relations. The GUSTO-I trial was designed to compare new thrombolytic strategies with standard thrombolytic regimens in the treatment of acute myocardial infarction.[6] The present study evaluates whether the extend and duration of myocardial ischaemia as measured by continuous ST-monitoring are associated with infarct size and left ventricular (LV) function in patients with acute myocardial infarction treated with thrombolytic therapy. Subsequently, the added value of continuous ST-monitoring to other patient characteristics for assessment of infarct size and LV function is evaluated.

Patients and Methods

Patients

The multicenter GUSTO ECG-monitoring substudy[13,14] included patients from the GUSTO angiographic substudy[7] and patients enrolled in the non-invasive part of the main study[6], as described previously. In brief, patients were eligible for GUSTO if they had chest pain lasting at least 20 minutes, up to 6 hours after symptoms onset and ST segment elevation at 60 milliseconds after the J-point (J+60 msec) ≥ 0.1 mV in two or more limb leads, or ST J+60 msec ≥ 0.2 mV in two or more precordial leads.[6] After informed consent, patients were randomized to one of four thrombolytic regimens: 1) streptokinase with subcutaneous heparin 2) streptokinase with intravenous heparin 3) accelerated alteplase (tPA) with intravenous heparin 4) combination with tPA, streptokinase and intravenous heparin. All patients received aspirin. In 10 out of 15 hospitals participating in the ECG monitoring substudy in Europe, patients were also enrolled in the angiographic substudy and were randomly assigned to coronary angiography at 90 minutes, 180 minutes, 24 hours or one week after start of thrombolytic therapy.[7] Patients assigned to 90 minutes angiography also underwent follow-up angiography one week later. The study design and technical considerations of the ECG-monitoring substudy[13,14] and the enzyme substudy[15] have been published recently.

Figure 3.2.1 Example of an overall ST-trend (= ST-segment changes over time) obtained from continuous ST-segment monitoring. The ST-trend characteristics (see text) evaluated in the present study to predict infarct size and left ventricular function can be obtained from the figure.



Admis ECG = ECG on admission; peak = peak ST-level; 50% = moment of 50% ST-reduction from the preceding peak ST-level; R1 and R2 are two recurrent ischemic episodes; μV = micro volt; the grey zones reflect the estimated area under the ST-trend.

The current analysis included only patients from the ECG-monitoring substudy, who were monitored by a vector-derived 12-lead ECG recording system (MIDA 1000, Ortivus Medical, Täby, Sweden), and in whom either infarct size or left ventricular function were determined.[13]

Enzymatic infarct size

Enzymatic infarct size was defined as the cumulative release of alpha hydroxybutyrate dehydrogenase (HBDH) activity per litre of plasma over 72 hours since the onset of symptoms, indicated as Q(72). The Q(72) value was divided by the normal HBDH content of human myocardium determined with the same assay, that is, 123 U per gram net weight, to obtain infarct size in gram-equivalents of myocardium per liter of plasma.[15]

Left ventricular ejection fraction

Left ventricular ejection fraction (LVEF) was measured from left ventricular angiograms by experienced angiographers without knowledge of previous obtained information from patient history, physical examination, electrocardiography and enzymatic infarct size.[7] For 123 patients the LVEF measured at the one week angiogram was used. In 22 patients this one week LVEF measurement was missing and the 90 minutes angiographic measurement was used instead. Since the ejection fraction did not change significantly between 90 minutes and one week after treatment initiation[7], this seemed justified.

Continuous ST-segment monitoring

Continuous ST-segment monitoring was performed for at least 18 hours from start of thrombolytic therapy. For each patient, the single lead with the most extensive ST elevation at J+60 msec was used to produce an overall ST-trend.[13,14] Potential ST-trend characteristics (figure 3.2.1) that were thought to be associated with enzymatic infarct size and LVEF were:

1. Peak ST level (Indicator of the extend of the ischaemia).
2. time (in minutes) to 50% ST-recovery from the preceding peak ST level since start of thrombolytic therapy (indicator of the time of reperfusion).

3. the area under the ST-trend curve from the onset of ST-monitoring to the moment of 50% ST-recovery (combined indicator of the extend of the ischemic area and the duration of coronary occlusion). The ST-area was estimated using the trapezoidal rule, by integration of the ST-trend until the zero ST-level (figure 3.2.1) and expressed in millivolt times minutes (mV.min). It reflected both the extend of the ST-elevation and the duration of the time to 50% ST-recovery.
4. the number of recurrent ischemic episodes following the 50% ST-recovery, defined as ST reelevation of ≥ 100 μV developed within a 10 minute period and lasting ≥ 60 seconds (indicator of stability or rather instability of reperfusion).
5. the total duration of all recurrent ST-elevation episodes together (a measure of total duration of reischaemia).
6. the sum of the area of all recurrent ST-elevation episodes, relative to the baseline ST-level (figure 3.2.1) before start of ST-reelevation (mV.min). This area reflected both the extend of ST reelevation and the duration of the recurrent ischemic episode(s).

If the ST-elevation on the ECG on admission was the highest measured ST deviation, this was taken as the peak ST-level. If the 50% ST-recovery had occurred before or at the start of monitoring, the time to 50% ST-recovery was the time interval between start of thrombolysis and start of monitoring. As a consequence, the area until 50% ST-recovery was estimated as zero. Similarly, if no recurrent ischemic episodes occurred, the number, total duration and sum of the area of these episodes was estimated as zero.

It has been demonstrated that patient characteristics documented at admission, such as the presence of anterior infarction and the time between onset of symptoms and start of treatment, predict both infarct size as well as the limitation of infarct size by thrombolytic therapy.[1,2,4,16,17,18] Therefore, we analyzed the value of ST-segment monitoring in the assessment of both outcomes, when added to determinants obtained from patient history (e.g. age, sex, previous MI and time from onset of symptoms to treatment initiation), physical examination (e.g. heart rate, Killip class), and findings on a 12-lead ECG recorded on admission like location of the infarct and the extend of ST-segment elevation.

Statistical analysis

Before start of the analyses it was appreciated that the sample size would not allow for detection of differences between infarct size or LVEF among the four GUSTO treatment groups. In GUSTO, both initial (90 minutes) patency and survival were significantly better in patients treated with accelerated tPA.[6,7] Moreover, in this subgroup early coronary patency and infarct size measurements were almost identical in the two tPA groups.[15] Accordingly, in the current analysis we decided to combine the two tPA groups (in deviation from the main GUSTO analysis) and to combine the two streptokinase regimens.

Data analysis was performed with standard software packages (SAS Institute Inc., Cary, release 6-08). Frequencies and median values of LVEF, infarct size and all potential predictors were calculated. An univariable linear regression analysis was used to select the significant determinants of enzymatic infarct size and LVEF among all potential predictors. Multivariable linear regression analysis was used to evaluate the most important predictors of infarct size and LVEF from continuous ST-monitoring. Subsequently, the characteristics of patient history and physical examination which were clinically relevant and significantly related to infarct size in the univariable analysis were included in a multivariable linear regression model. To evaluate the added value of ST-monitoring to assess both outcomes, we included all ST-monitoring findings to this model. To adjust for the possible treatment effect on infarct size, the indicator for thrombolytic therapy was included in every model irrespective of its significance level. The same procedure was followed for prediction of LVEF.

Results

Of the 406 patients who were monitored by the vector-derived 12-lead ECG recording system in the GUSTO ECG-monitoring substudy, 46 were excluded either because the continuous ECG-monitoring was started more than 60 minutes after initiation of thrombolytic therapy or therapy was started more than six hours since onset of symptoms. Ninety one patients were excluded because of technical failures or missing ECG trend data.[13] Thus, 269 patients had vector-derived 12-lead ECG recordings suitable for analyses. Table 3.2.1 shows the baseline characteristics of these patients which did not differ from the 41,021 patients in the main GUSTO-I trial.[6] Of the 269 patients that underwent ST-monitoring, 206 patients also participated in the enzyme substudy. The mean Q(72) of these patients was 4.7 g-eq/l. Of the 269 patients, 231 participated in the angiographic substudy of whom 214 patients underwent angiography with a mean LVEF of 58%. These 206 and 214 patients were used for prediction of infarct size and LVEF, respectively. In 155 patients both outcomes were measured.

Table 3.2.1 Median values and frequencies of the most important characteristics from patient history and physical examination (N = 269).

	Median or %
age (year)	61 (52-69)
females (%)	0
Previous myocardial infarction (%)	14
Previous angina (%)	47
Heart rate (beats/minutes)	72 (60-81)
Anterior infarct location (%)	42
ST elevation on admission (millivolt)	0.4 (0.2-0.6)
Time to treatment (hours)	169 (125-220)
Alteplase therapy use* (%)	48
Killip class III or IV (%)	1

N = Number; LV = left ventricular; g-eq = gram-equivalents.
Numbers between parentheses are the 25th and 75th percentiles.
* Either accelerated tPA or combined tPA and streptokinase.

Table 3.2.2 shows the descriptive statistics for the ST-segment monitoring characteristics. The median peak ST-level and time to 50% ST-recovery were 0.5 mV and 46 minutes, respectively. The area under the ST-trend until 50% ST-recovery ranged from 0, i.e. ST-recovery before start monitoring (52 patients) to 246 mV.min with a median of 12 mV.min. Recurrent ischemic episodes were present in 97 patients (36%), 22 (8%) had more than 2 ischemic episodes of whom two patients had 14 and one patient had 40 episodes. The total duration of all recurrent ischemic episodes ranged from 2 minutes to 8 hours with a median of 16 minutes. The area under all recurrent ischemic episodes varied between 0.2 and 111 mV.min with a median of 2 mV.min.

Univariable analysis

Table 3.2.2 Median values and frequencies of the characteristics of continuous ST-segment monitoring.

ST-segment monitoring characteristics	N	Median or %
Peak ST-level (mV)	269	0.5 (0.3-0.7)
Time to 50% ST-recovery (min)	269	46 (24-84)
AUC until 50% ST-recovery (mV.min)	269	11 (1-28)
Number ischemic episodes	269	
0 (%)	172	64
1 (%)	61	23
2 (%)	14	5
3 - 40 (%)	22	8
Total duration of all reischemic episodes (min)	97	16 (7-44)
AUC of all reischemic episodes (mV.min)	97	2 (1-8)

N = Number; min = minutes; AUC = area under the ST-trend curve; mV = millivolt.
Numbers between parentheses are the 25th and 75th percentiles.

Previous MI, previous angina, time to treatment, heart rate, infarct location, and ST elevation on admission were all associated with infarct size and (except time to treatment) LVEF (table 3.2.3). LVEF was also associated with age and sex. Although there was a small effect of tPA therapy (accelerated tPA or tPA in combination with streptokinase) on infarct size and LVEF, the 95% CI of the regression coefficients were very wide. Similar results were found, if accelerated tPA with heparin was compared to the three other regimens as was done in the main GUSTO trial.[6]

Both a higher peak ST-level, a longer duration to 50% ST-recovery and the combination of both as reflected by the area under the curve until 50% ST-recovery, showed a significantly larger infarct size and lower LVEF. If we excluded the 52 patients with an ST-recovery before start of ST-monitoring these associations did not change. The infarct size was 1.2 g-eq/l higher and the LVEF 1.3% lower in patients with recurrent ischemic episodes though the 95% CI of these differences were very wide. However, in patients with recurrent ischemic episodes, the infarct size and LVEF significantly

Table 3.2.3 Regression coefficients and 95% confidence intervals (95% CI) of univariable linear regression models predicting enzymatic infarct size (Q(72) in g-eq/l) and left ventricular ejection fraction (LVEF in %).

	Q(72) (n=206)	LVEF (n=214)
	coefficient (95% CI)	coefficient (95% CI)
Patient characteristics		
Age (per 10 years)	0.1 (-0.3; 0.5)	-2 (- 4; 0)
Female	-0.2 (-1.3; 0.9)	5 (0; 10)
Previous MI	-1.2 (-2.5; 0.1)	-12 (-18;-6)
Previous angina	-1.1 (-2.0;-0.2)	-4 (- 8; 0)
Time to treatment (per hour)	0.3 (-0.1; 0.7)	-0.2 (-27;27)
Heart rate (per 10 beats/min)	0.3 (0.1; 0.5)	-3 (- 4;-2)
Anterior infarct	1.0 (0.1; 1.9)	-10 (-14;-6)
ST elevation on admission (per mV)	4.4 (2.8; 6.0)	-7 (-14; 0)
Alteplase therapy	-0.6 (-1.6; 0.6)	1 (- 3; 5)
ST-monitoring characteristics		
Peak ST-level (per mV)	4.3 (2.8; 5.8)	-6.9 (-14; 0)
Time to 50% ST-recovery (per 30 minutes)	0.2 (0.0; 0.4)	-0.9 (-1.7;-0.1)
AUC until 50% ST-recovery (per 10 mV.min)	0.3 (0.1; 0.5)	-0.8 (-1.4;-0.2)
Presence of ischemic episodes	1.2 (0.3; 2.1)	-1.3 (-5.5; 2.9)
Number of reischemic episodes (per episode)	0.2 (0.0; 0.4)	-1.4 (-2.8; 0)
Total duration reischemic episodes (per 10 minutes)	0.1 (0.0; 0.2)	-0.5 (-1.1; 0.1)
AUC of all reischemic episodes (per 10 mV.min)	0.6 (0.1; 1.1)	-2.0 (-4.2; 0.2)

G-eq/l = gram-equivalents per liter; n= number; min = minutes; mV = millivolt; AUC = area under the ST-trend curve.

increased and decreased, respectively, with a greater number of episodes, a longer total duration of the episodes, and a greater area under the curve. Other variables derived from continuous ST-segment monitoring were not associated with either outcome.

Multivariable analysis

The peak ST-level, area until 50% ST-recovery and area under the recurrent ischemic episodes were the independent ST-trend predictors of infarct size (table 3.2.4). The area under the ST-trend till 50% ST-recovery was the only independent ST-trend predictor of LVEF (table 3.2.4). The number of episodes was stronger associated with infarct size and with LVEF than the total duration of the ischemic episodes. Compared to the number, the area under all episodes was stronger associated with both outcomes (data not shown).

Table 3.2.4 Mutually adjusted regression coefficients of the ST-monitoring factors predicting infarct size (Q(72) in g-eq/l) and left ventricular ejection fraction (LVEF in %).

	Q(72)	LVEF
Peak ST-level (per mV)	3.8 (5.4;2.2)*	-3.4 (-10.8;4.0)
AUC until 50% ST-recovery (per 10 mV.min)	0.1 (0.0;0.2)	-0.8 (-1.4;-0.2)
AUC of reischemic episodes (per 10 mV.min)	0.5 (0.1;0.9)	-2 (-5;1)

G-eq/l = gram-equivalents per liter; AUC: area under the ST-trend curve; min = minutes; mV = millivolt.

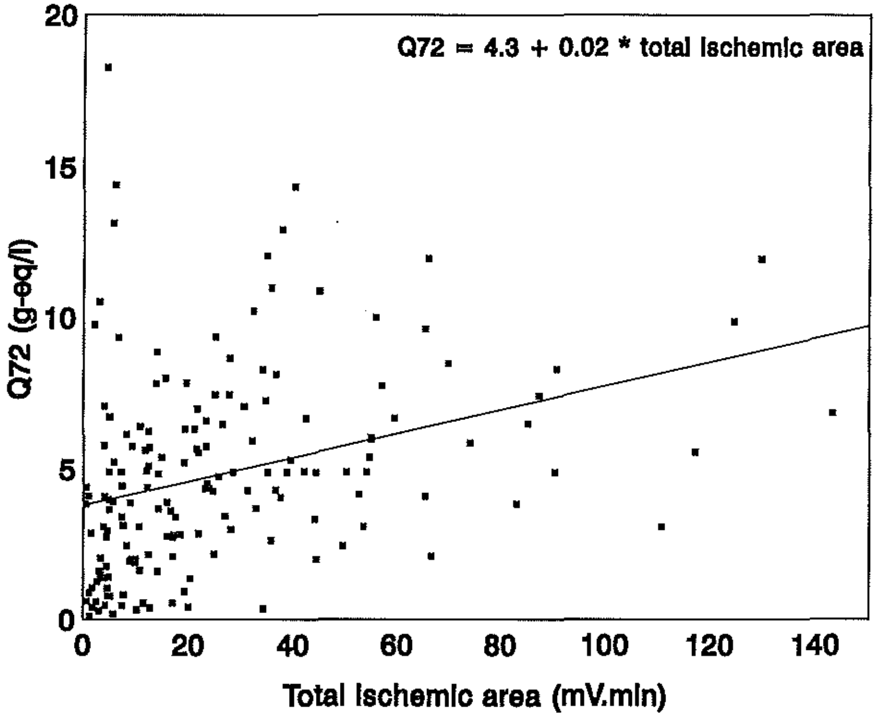
* The numbers between parentheses is the 95% confidence interval.

Similar results were found after excluding the exceptional patient with 40 episodes. As more than two third of the patients had no ischemic episodes (corresponding area estimated as zero) which might had influenced the estimated regression coefficient, the two ST-trend areas were combined to a summed area. This area reflected the extend and duration of the total ischaemia in the patient in 24 hours and it was very strongly associated with both outcomes. These associations did not change after excluding the patients with 50% ST-recovery before start of ST-monitoring (figures 3.2.2 and 3.2.3).

Time to treatment and ST elevation on admission were independent predictors of infarct size (model R^2 = 17%). After addition of all ST-monitoring parameters, treatment delay, ST-elevation on admission and the total ischemic area were the independent predictors (model R^2 = 23%, table 3.2.5). Although both provided largely the same information, ST-level on admission was stronger associated compared to peak ST-level. Excluding 15 patients who underwent early angioplasty which could have influenced the time to 50% ST-recovery, the occurrence of reischaemia and the infarct size, did not change the results.

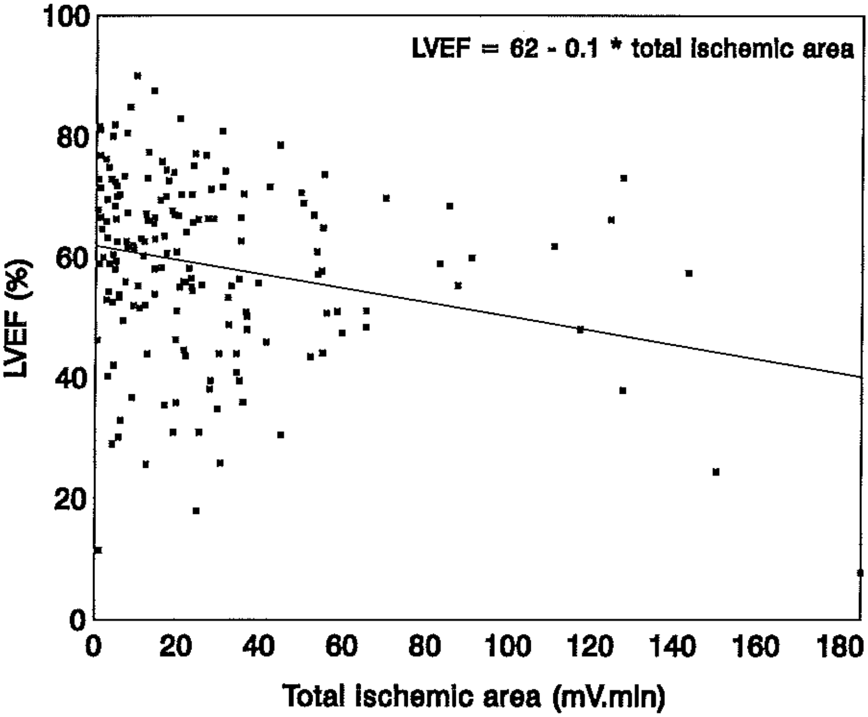
Previous MI, anterior MI location, heart rate and age were independent determinants of LVEF (model R^2 = 27%). These variables remained significant when the ST-monitoring characteristics were added to the model, of which total ischemic area under the ST-trend was the only independent predictor (model R^2 = 32%, table 3.2.6).

Figure 3.2.2 Association of the total ischemic area (= area until 50% ST-recovery + area under all recurrent ischemic episodes) with infarct size (Q(72)) for all patients of whom the moment of 50% ST-recovery occurred after start monitoring.



G-eq/l = gram-equivalents per liter; mV.min = millivolt times minutes.

Figure 3.2.3 Association of the total ischemic area (= area until 50% ST-recovery + area under all recurrent ischemic episodes) with left ventricular ejection fraction (LVEF) for all patients of whom the moment of 50% ST-recovery occurred after start monitoring.



mV.min = millivolt times minutes.

In this analysis, 22 patients had 90 minutes and 46 patients had 180 minutes angiography without a one week angiogram. Their LVEF was obtained long before the ST-monitoring was ended. Because this might confound the relationship between ST-trend analysis and LVEF a second analysis was performed without these patients. This did not change the results in the association with LVEF. Similarly, excluding the 15 patients who underwent early angioplasty did not change the results.

Table 3.2.5 Mutually adjusted regression coefficients of patient history, physical examination and continuous ST-monitoring, predicting infarct size (Q(72) in g-eq/l).

Patient history + physical examination + ST-monitoring	Q(72)
Intercept	1.7
Alteplase	-0.9 (-2.0;0.1)*
Time to treatment (per hour)	0.4 (0;0.8)
ST elevation on admission (per mV)	4.6 (3.0;6.2)
Total ischemic area (per 10 mV.minutes)†	0.2 (0.1;0.3)
R^2	23%

G-eq/l = gram-equivalents per liter; mV = millivolt; AUC = area under the ST-trend curve.

* The numbers between parentheses give the 95% confidence interval.

† Sum of the area under the ST-trend until 50% ST-recovery and the area under all recurrent ischemic episodes.

Discussion

This analysis from GUSTO-I evaluated the association of continuous ST-monitoring characteristics and infarct size and left ventricular function in patients with acute myocardial infarction treated with thrombolytic therapy. The area under the ST-trend curve until 50% ST-recovery and the sum of the area under the recurrent ischemic episodes (or a combination of both) appeared to be predictors of infarct size and LVEF, independent from other patient characteristics. This supports the physiologic hypothesis that both the extend and duration of myocardial ischaemia (both included in the estimated area under the ST-trend curve) determine the myocardial damage and, thus, may predict the infarct size and ejection fraction.

Because the true baseline of the ST-segment of the GUSTO patients, i.e. before the onset of myocardial ischaemia, was not known, the area to the moment of 50% ST-recovery was estimated relative to the zero ST-level. In addition, a subgroup analysis was done among patients with recurrent ischemic episodes in which both the area until 50% ST-recovery and the sum area of the ischemic episodes were estimated relative to a baseline ST-level. This baseline ST-level was estimated by the average of all baseline ST-

Table 3.2.6 Mutually adjusted regression coefficients of patient history, physical examination and continuous ST-monitoring, predicting LVEF (in %).

Patient history + physical examination + ST-monitoring	LVEF
Intercept	91
Alteplase	2 (- 2; 6)*
Previous MI	-12 (-17;-7)
Anterior location	-8 (-12;-4)
Age (per 10 years)	-2 (- 4; 0)
Heart rate (per 10 beats/min)	-2 (- 3;-1)
Total ischemic area (per 10 mV.minute)†	-1.0 (-1.5;-0.5)
R^2	32%

LVEF = left ventricular ejection fraction; mV = millivolt; AUC = area under the ST-trend curve.

* The numbers between parentheses give the 95% confidence interval.

† Sum of the area under the ST-trend until 50% ST-recovery and the area under all recurrent ischemic episodes.

levels after 50% ST-recovery as measured prior to each reischemic episode. This restricted analysis provided similar results.

The present study used objective ST-recovery criteria as suggested previously.[8-12] These studies have demonstrated that ST-trend characteristics reflect patency of the infarct-related artery. Previous GUSTO-I analyses have shown that early patency of the infarct-related artery (within 90 minutes) by thrombolytic therapy reduces infarct size and improves left ventricular function and that 72% to 74% of the patients were patent (TIMI score 2 and 3) within 90 minutes.[7,13,15] Similarly, in our study, 75% of the patients have 50% ST-recovery within 84 minutes. Therefore, the present study also supports the hypothesis that ST-monitoring characteristics reflect coronary patency. Furthermore, the present study can be regarded as an addition to the early analyses of GUSTO-I data in which simplified criteria for ST-recovery (the presence or absence of 50% ST-recovery and ST-reelevations) were evaluated to predict patency of the infarct-related artery.[13] We studied whether the extend, speed, (in)stability and duration of ST-recovery and, therefore, of reperfusion, provide information about (the thrombolytic effect on) infarct size and left ventricular function, directly. To our knowledge, this has not been evaluated before. Combining the results of the previous[13] and the present study, it is obvious that continuous ECG monitoring may diagnose patient subgroups without apparent reperfusion which, presumably, develop a larger infarct size and worse left ventricular function. These subgroups may benefit from additional therapy. Hence, the clinical relevance of ST-monitoring directs to the tailoring of thrombolytic therapy.[1,19]

The limitations of the GUSTO-I ECG monitoring substudy as described previously[13] also apply to the present study. First, the participating centers in GUSTO-I were relatively inexperienced in applying continuous ECG monitoring systems in the setting of acute myocardial infarction. This could partly be the reason why more than hundred patients had to be excluded because of a delayed start of recording, missing trend data or technical failures. Since these reasons for the missing and uninterpretable data were unlikely to be related to infarct size or ejection fraction, and neither to the potentially observable ST-monitoring result, we believe that it will not have biased our results. Second, for patients randomized to the angio group of 90 and 180 minutes, the ECG recordings had gaps in the ST-trend data during the procedure in the angiography room. As changes of vessel status as well as in the ECG may occur rapidly, the ST-segment could have changed during that period, e.g. peak ST-level, 50% ST-recovery or recurrent ischaemia could have occurred. The ST-trend data during that period had to be extrapolated from the slope of ST-trend data before and after the angiography procedure. This may have resulted in less accurate estimates of the areas under the ST-trend. The same applies to situations when the recording system had to be disconnected during transportation from the coronary care unit to the angiography room. Also, the influence of the time intervals between start of thrombolytic treatment and start of ST-monitoring, and between onset of symptoms and treatment initiation on the ST-monitoring characteristics as discussed in that study apply in a similar way to the present study.

With respect to determinants of infarct size and ejection fraction other than ST-trend characteristics, our findings are in agreement with previous large studies.[1,2,4,16-18] These studies also have found that ejection fraction is lower in older patients, with anterior infarction and previous myocardial infarction and infarct size is larger if treatment delay is longer and if ST-elevation on admission is higher. As compared to these large studies, the relatively small number of patients is the most plausible reason for not finding a significant association between, for example, infarct location and previous myocardial infarction with infarct size, and treatment delay and ST-elevation on admission with ejection fraction.

In conclusion, the area under the ST-trend till 50% ST-recovery and the (sum of the) area of recurrent ischemic episode(s), which reflect both the extend and duration of myocardial ischaemia, independently predict infarct size and left ventricular function, rather than the peak ST-level, time to 50% ST-recovery, the number and duration of recurrent ischemic episodes.

References

1. Arnold AER, Jaegere P de, Schröder R, Brüggeman T, Simoons ML, Lubsen J. Expected infarct size without thrombolysis: a concept that helps to select patients with evolving myocardial infarction for thrombolytic therapy. In: Arnold AER. Benefits and risks of thrombolysis for acute myocardial infarction. Rotterdam: Erasmus University 1990. Thesis.

2. Gruppo Italiano per lo studio della streptochinasi neel'infarto miocardico (GISSI): Long term effects of intravenous thrombolysis in acute myocardial infarction: final report of the GISSI study. Lancet 1987;1:871-4.

3. Werf F van der, Arnold AER. Intravenous tissue plasminogen activator and size of infarct, left ventricular function, and survival in acute myocardial infarction. Br Med J 1988;297:1374-9.

4. ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction. Lancet 1988;2:349-60.

5. Simoons ML, Vos J, Tijssen JGP, et al. Long term benefit of early thrombolytic therapy in patients with acute myocardial infarction: 5 year follow-up of a trial conducted by the Interuniversity Cardiology Institute of The Netherlands. J Am Coll Cardiol 1989;14:1609-15.

6. The GUSTO investigators. An international trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993;329:673-82.

7. The GUSTO Angiographic Investigators. The effects of tissue plasminogen activator, streptokinase, or both on coronary-artery patency, ventricular function, and survival after acute myocardial infarction. N Engl J Med 1993;329:1615-22.

8. Krucoff MW, Green CE, Satler LF, et al. Non-invasive detection of coronary artery patency using continuous ST-segment monitoring. Am J Cardiol 1986;57:916-23.

9. Hogg KJ, Hornung RS, Howie CA, Hockings N, Dunn FG, Hillis WS. Electrocardiographic prediction of coronary artery patency after thrombolytic treatment in acute myocardial infarction: use of the ST-segment as a non-invasive marker. Br Heart J 1988;60:275-80.

10. Dellborg M, Riha M, Swedberg K. Dynamic QRS-complex and ST-segment monitoring in acute myocardial infarction during recombinant tissue-type plasminogen activator therapy. Am J Cardiol 1991;67:343-9.

11. Hohnloser SH, Zabel M, Kasper W, Meinertz T, Just H. Assessment of coronary patency after thrombolytic therapy: accurate prediction utilizing the combined analysis of three noninvasive markers. J Am Coll Cardiol 1991;18:44-9.

12. Krucoff MW, Croll MA, Pope JE, et al. Continuously updated 12-lead ST-segment recovery analysis for myocardial infarct artery patency assessment and its correlation with multiple simultaneous early angiographic observation. Am J Cardiol 1993;71:145-51.

13. Klootwijk P, Krucoff MW, Langer A, et al for the GUSTO ECG-ischaemia monitoring substudy. Non-invasive prediction of reperfusion and coronary patency by continuous ST-segment monitoring in the GUSTO trial. Eur Heart J 1996;17:689-98.

14. Krucoff MW, Green CL, Langer A, et al. Global utilization of streptokinase and tPA for occluded arteries (GUSTO) ECG-monitoring substudy. Study design and technical considerations. J Electrocardiol 1993;26 (suppl):249-55.

15. Baardman T, Hermens WT, Lenderink T, et al. Differential effects of tissue plasminogen activator and streptokinase on infarct size and on rate of enzyme release: influence of early infarct related artery patency. The GUSTO Enzyme Substudy. Eur Heart J 1996;17:237-46.

16. Bär FW, Vermeer F, Zwaan C de, et al. Value of admission electrocardiogram in predicting outcome of thrombolytic therapy in acute myocardial infarction. Am J Cardiol 1987;59:6-13.

17. Vermeer F, Simoons ML, Bär FW, et al. Which patients benefit most from early thrombolytic therapy with intracoronay streptokinase? Circulation 1986;74:1379-89.

18. Christian TF, Schwartz RS, Gibbons RJ. Determinants of infarct size in reperfusion therapy for acute myocardial infarction. Circulation 1992;86:81-90.

19. Simoons ML, Arnold AER. Tailored thrombolytic therapy: a perspective. Circulation 1993;88:2556-64.

Chapter 3.3

Efficiency optimisation of the selection period in therapeutic trials

Introduction

Before patients are selected for a therapeutic trial they have to fulfil all inclusion and exclusion criteria. Most of these criteria reflect the treatment indication whereas some are derived from logistic or ethical reasons. If immediate treatment is not required, measuring the inclusion and exclusion criteria can be extended over a certain period of time during which patients are repeatedly examined at certain time intervals. This results in a stepwise exclusion process which is called the qualification or selection period.[1] In this paper, consecutive examinations are referred to as "visits". A selection period comprising several visits rather than examining all criteria at once may be favourable for several reasons. First, to obtain an adequate assessment of certain characteristics. For example, blood pressure and serum cholesterol level[2-7] require more than one measurement at different times because of within person variability and regression towards the mean.[8-11] Second, a selection period may be important if no historical or medical data of the participants are available which may occur in trials conducted by investigators who are not the treating physicians. Third, spreading measurements over consecutive visits starting with the simple and non-burdening ones at an initial visit and continuing with the more expensive and invasive examinations at a later stage, may reduce the costs. Measuring all eligibility criteria at the very first visit may also result in more patient refusal.

This paper describes an approach for optimising the efficiency of a selection period in randomised trials, i.e. to obtain a maximum of randomisations with a minimum of examinations. The principles are illustrated using data from the selection period of a large trial on the efficacy of a cholesterol lowering drug.[12]

Theory

In a selection period, N_0 potential participants enter the first visit at which n_1 subjects are excluded and N_1 (= N_0-n_1) subjects are invited for the second visit. Similarly, of the N_1 participants reaching visit two, n_2 will be excluded and N_2, subjects will be examined at the third visit. This stepwise selection process continues until randomisation (R) and N_R subjects are included into the trial. The costs to randomise one subject can simply be estimated. If k ($1 \leq k \leq R$) is the number of visits passed by a subject and C are the measurement costs per visit, then $\Sigma^k_{i=1}$ C_i are the total costs for this subject. The mean costs per subject, c, is obtained by taking the average of this calculation over all subjects. The proportion randomised, p, is estimated by the number randomised divided by the total number that initially entered the selection period (N_R/N_0). The mean costs per randomisation is calculated by c/p and its standard error can be estimated using the delta method[13], based on the standard error of c and p and their correlation.

Prediction of exclusion at a later visit using information obtained at previous visits means that the subject will be excluded before the exclusion criterion is actually measured. This will decrease costs if the prediction is correct and will increase costs if the prediction is incorrect. In this predictive context, the outcome is defined as "excluded or not at a later visit by a certain criterion". The relative number excluded by a certain criterion, determines whether the criterion can serve as an outcome. Prediction of criteria with low exclusion frequencies will not generally lead to substantial increases in efficiency. Given a selected outcome, all variables measured at visits preceding the visit of the outcome are potential predictors.

Suppose a prediction model that is fit after visit i to predict exclusion by criterion E at visit k ($k > i$), and is applied to the (initial) selection period. The new selection strategy would be to withdraw each subject at visit i with an estimated probability of exclusion above a chosen cut-off value (the subject is called "positive"). Positive subjects contribute to the saving of the costs of all measurements between visit i and visit k. However, false positive predictions also lead to potential reduction in the number of randomisations. They could have been randomised unless they were excluded before visit k, by other criteria at visit k or by criteria measured between visit k and randomisation. The mean costs per randomisation in the new strategy, c^*/p^*, can be estimated in the same way as described above for the null situation, i.e. the initial strategy of the selection period without using the prediction model. If $c/p - c^*/p^*$ is significantly higher than zero, the new strategy is preferable. The standard error of this difference can be computed by the delta method, based on the standard errors of c, p, c^*, p^* and their correlations.[13]

Illustration

Design

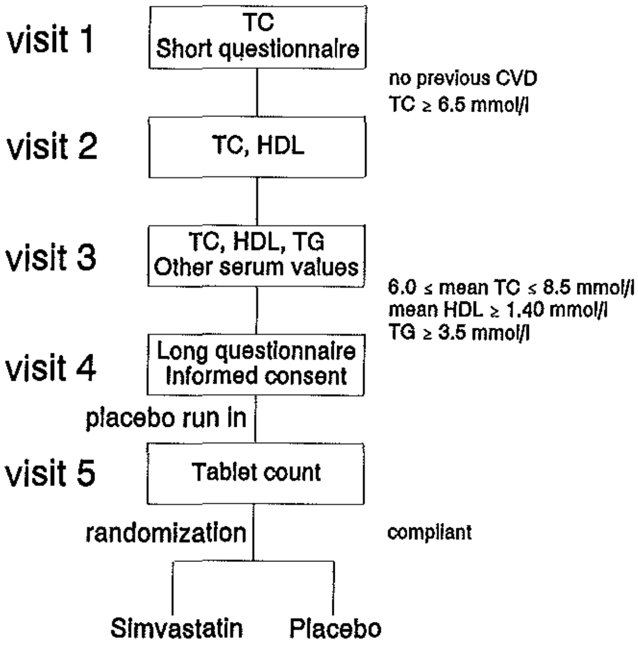
The theory above was evaluated using interim data from the selection period of the Rotterdam Cardiovascular Risk Intervention trial (ROCARI).[12] ROCARI was designed as a randomised, placebo-controlled, primary prevention study to evaluate the effect of a cholesterol-lowering drug (simvastatin) on fatal and non-fatal atherosclerotic coronary heart diseases (CHD) in 9,000 men, aged between 40 and 70 years. All participants had primary hypercholesterolemia without clinical evidence of CHD. Subjects from the Dutch city of Rotterdam and suburbs entered the selection period either by general practitioner referral or through a direct mail procedure. The selection period comprised a stepwise exclusion process over five consecutive visits at one month intervals (Figure 3.3.1). Participants could refuse further participation at any time.

At the first visit, subjects were eligible for further screening if their serum total cholesterol level was 6.5 mmol/l or over and if they satisfied additional criteria checked by a short questionnaire on medical and family history. At the second visit, serum total cholesterol and high density lipid cholesterol (HDL) levels were measured. There were no criteria to proceed to the third visit during which the latter two measurements were repeated. Total cholesterol and HDL at visit two were averaged with the values measured at visit three. At visit three, triglycerides and other serum measurements were performed. If the mean total cholesterol level was in the range of 6.0-8.5 mmol/l, the mean HDL level below 1.40 mmol/l, the triglyceride level below 3.5 mmol/l and the other biochemical values were in the normal range, participants were invited for the fourth visit. After the subjects satisfied a large number of criteria at the fourth visit and gave written informed consent, they entered a placebo run-in phase. At visit five, after satisfying a final evaluation of all criteria, the subjects were randomised on a 1:1 basis to either simvastatin or placebo. At time of the analysis 30,372 subjects had entered the selection period at visit one and over 1,100 had been randomised. The present analysis is based on data of a cohort of 6,544 men that completed the selection period, unless they were excluded before they could be randomised.

Methods

For the present analysis, criteria leading to the exclusion of 50 subjects or over were selected as outcomes. Given a selected outcome on a particular visit, we constructed multivariable logistic regression models including all variables obtained from previous visits using data of the first 2,200 subjects that entered the ROCARI selection period (referred to as the derivation set).[14] Continuous variables were included as continuous terms only if a linear relation was plausible. The multivariable models were constructed in concordance with the chronological order in which data became available during the selection period. The initial model comprised all variables that were measured at visit one. Model reduction was based on clinical relevance and significance level. We compared the information content of various reduced models with the initial model using the area under the Receiver Operating Characteristic curve (ROC area) and its standard error; both were estimated using the non-parametric approach.[15,16]

Figure 3.3.1 Flow chart of the five visits of the ROCARI selection period (see text).



CVD = cardiovascular diseases; HDL = high density lipid cholesterol; TC = total cholesterol; TG = triglycerides. Mean TC and mean HDL are the average TC and HDL levels of the measurements at visit 2 and 3.

Model reduction was done to obtain an efficient model, i.e. the model with a minimum number of determinants that did not have a significantly lower ROC area than the initial model. Subsequently, for each patient the probability of being excluded by the outcome was estimated by the fitted model. We estimated the costs per randomisation for nine equidistant cut-off values on the scale of estimated probabilities. The costs per randomisation described an optimum curve over all cut-off values because the number of true and false predicted exclusions decreased unequally with increasing cut-off value. Consequently, the difference in costs per randomisation between the initial and the new selection strategy which uses the prediction model also described an optimum curve over all cut-off values. This optimum corresponded to the most efficient probability threshold of the model. When the optimum was lower than zero, the outcome could not efficiently be predicted and the model was rejected.

The same approach was applied to all variables measured at the second visit after they were added to the fitted model from visit one. This yielded the fitted model of visit two. This procedure was chronologically pursued for each subsequent visit until the visit on which the outcome itself was determined. The entire approach was applied to all selected outcomes.

We applied the optimum probability thresholds of the fitted "visit-models" to data of 4,344 subjects that subsequently entered the selection period. The difference in costs per randomisation between the initial and the new selection strategy with application of the prediction model was estimated.

Results

Table 3.3.1 shows the number of patients screened and excluded per visit for the cohort of 6,544 men. From table 3.3.1 the following exclusion criteria were selected as an outcome at visit three: mean HDL of 1.40 mmol/l or higher (high HDL), mean total cholesterol lower than 6.00 mmol/l (low TC), mean total cholesterol higher than 8.50 mmol/l (high TC) and fasting triglycerides of 3.50 mmol/l or higher (high TG). Because the total number excluded at visit four and five of ROCARI was still too small at the time of analysis, no single criterion was selected for these visits. Willingness to give informed consent at visit four was not selected because association with previous data was unlikely.

Table 3.3.1 The number examined and excluded on the main exclusion criteria per visit for a cohort of 6,544 men that could have completed the selection period of ROCARI.

	visit 1	visit 2	visit 3	visit 4	visit 5
Measurement costs per visit (US $)	90	63	126	207	144
Number examined	6544	2378	2186	1261	1087
Number proceeding to next visit	2378	2186	1261	1087	985
Number excluded*	4166	152	925	174	102
TC < 6.5 mmol/l	3722	7	0	1	-
Mean TC(2+3) < 6.00 mmol/l (low TC)	-	-	215	-	1
Mean TC(2+3) > 8.50 mmol/l (high TC)	-	-	78	-	-
Mean HDL(2+3) ≥ 1.40 mmol/l (high HDL)	-	-	366	1	1
Fasting triglycerides ≥ 3.5 mmol/l (high TG)	-	-	167	2	1
No informed consent	-	-	-	79	1
Tablet non-compliance during run-in	-	-	-	3	30
Clinical evidence previous CVD	183	30	23	38	22
Use other lipid lowering drugs	130	20	6	-	-
Non-lipids out of range	-	-	14	7	21
Unable or refusal for further participation	233	128	105	34	13
Sum of 15 other criteria	65	6	137	32	38

TC = Total cholesterol; CVD = Cardiovascular diseases; TG = Triglycerides.

* The sum of number excluded per criterion exceeds the number excluded per visit due to the possibility of more than one reason for exclusion per participant.

For each selected outcome table 3.3.2 shows the ROC area of the initial and corresponding reduced model of visit one and two. The reduced model of visit one to predict high HDL included body mass index and age. The reduced model of visit one to predict low TC included total cholesterol level and smoking history. To predict high TC this model included total cholesterol level, smoking history and the presence of an elevated cholesterol in the past. These determinants also remained in the reduced

Table 3.3.2 Area under the ROC curve of the initial and reduced prediction model per visit for the four selected outcomes at visit three based on data obtained from the derivation set (N=2,200).

			Visit 1		Visit 2	
Outcome	Ntot	Nexcl	Initial model	Reduced model	Initial model	Reduced model
High HDL	721	98	0.65 (0.58-0.72)*	0.63 (0.57-0.70)	0.98 (0.97-0.99)	0.98 (0.97-0.99)
Low TC	721	68	0.73 (0.64-0.82)	0.71 (0.62-0.80)	0.95 (0.91-1.00)	0.95 (0.91-0.99)
High TC	721	31	0.92 (0.83-1.00)	0.90 (0.78-1.00)	0.97 (0.89-1.00)	0.96 (0.88-1.00)
High TG	721	57	0.73 (0.63-0.83)	0.69 (0.58-0.80)	0.76 (0.66-0.86)	0.76 (0.66-0.86)

Ntot = Total number on which the models were based (i.e. the number by whom the outcome was measured); Nexcl = Number excluded by the outcome; TC = Total cholesterol; TG = Triglycerides.
* Numbers within parentheses is the 95% confidence interval.

model of visit one to predict high TG. However, only exclusion by high TC could adequately be predicted at visit one (ROC area of the reduced model was 0.90). The ROC area of the initial model at visit two for prediction of high HDL, low TC, and high TC was 0.98, 0.95 and 0.97, respectively. This initial model included the corresponding reduced model of visit one plus the total cholesterol level and HDL level measured at visit two. However, to predict high HDL the reduced model with HDL level measured at visit two only, had the same ROC area as the initial model. Similarly, exclusion by low TC as well as by high TC could equally be predicted by the model including total cholesterol level measured at visit two only.

In ROCARI, the costs per randomisation in the initial selection strategy were US $ 1,444. Figure 3.3.2 shows the distribution of the difference in costs per randomisation between the initial selection strategy and after application of a new strategy. This new strategy used the reduced model derived at visit two to predict exclusion by high HDL at visit three. The optimum was between 0.8 and 0.95 and defined at 0.9. Excluding everyone with an estimated probability higher than 0.9 would save more than US $ 30 per randomisation. For probability thresholds above 0.9, the number of true predicted exclusions decreased more than the number of false predicted exclusions, resulting in a lower difference in costs per randomisation. This also occurred for probability thresholds lower than 0.9. Here, the number of false predicted exclusions increased more compared to the number of true predicted exclusions. For thresholds below 0.4, the loss of randomisations by the false predictions even outweighed the saved costs from true predictions. The result was an increase of costs per randomisation.

The fitted model of visit one to predict high HDL could not decrease the costs per randomisation. For the prediction of exclusion by low TC, the optimum probability threshold of the reduced model at visit one and two was 0.4 and 0.7, respectively, whereas for the prediction of high TC these thresholds were 0.6 and 0.5, respectively. Although the ROC areas of the two fitted models to predict triglyceride levels of 3.50 or higher were not extremely low (table 3.3.2), the costs per randomisation could not be decreased for either model.

The reduced models of visit two to predict high HDL, low TC, and high TC comprised just one determinant. For high HDL this was the HDL level measured at visit two whereas for low TC and high TC this was the total cholesterol level measured at visit two. Therefore, the estimated probability thresholds of these reduced models could directly be presented as a HDL level or as total cholesterol levels, respectively. For example, to withdraw everyone with blood HDL levels over 1.48 mmol/l at visit two was similar to applying the derived model and using the probability threshold of 0.9 (figure 3.3.2). For the prediction of low TC and high TC, the total cholesterol levels of visit two could analogously be defined at 5.5 or lower and 8.8 or higher, respectively.

Figure 3.3.2 Distribution of the difference in costs per randomisation (in US $) between the initial selection strategy and the new strategy, over 10 cut-off values on the scale of estimated probabilities of the reduced model at visit two to predict exclusion by mean HDL ≥ 1.40 mmol/l at visit three. Analysis based on data obtained from the derivation set (N=2,200).

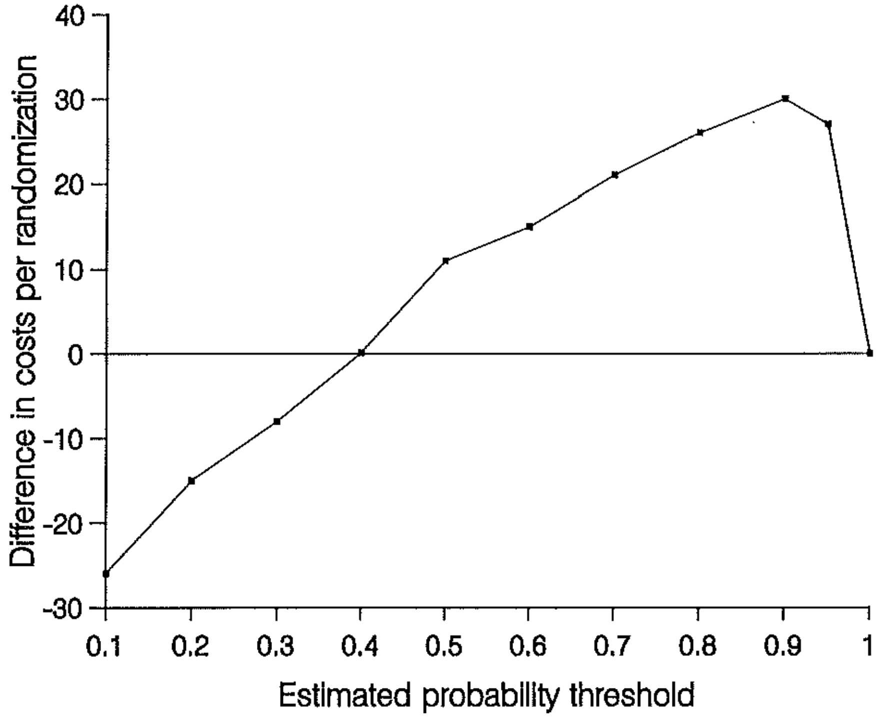


Table 3.3.3 summarizes the results after application of the optimum thresholds of the selected models to the 4,344 patients that subsequently entered the ROCARI selection period. The last column shows the saved costs over the remaining 8,657

Table 3.3.3 Estimated difference in costs per randomization and total saved costs, expressed in US $, if the reduced models of table 5.2 were applied to the ROCARI selection period, based on data of 4,344 men.

Outcome *Reduced model*	Excluded by the outcome[*]	Not excluded by the outcome[†]	Costs per randomizati on new strategy	Difference (95% CI)[‡]	Total saved costs[#] (95% CI) (US $ x 1000)
High HDL					
Model2: ≥ 0.9	186	9	1414	30 (20;40)	260 (173;346)
Low TC					
Model1: ≥ 0.4	3	0	1442	1 (-1; 3)	9 (-9; 26)
Model2: ≥ 0.7	83	13	1427	17 (11;23)	147 (95;199)
High TC					
Model1: ≥ 0.6	5	3	1443	1 (-3; 5)	9 (-26; 43)
Model2 : ≥ 0.5	36	7	1435	9 (5;13)	78 (43;113)
All optimal thresholds	333	26	1392	52 (39;65)	450 (338;563)

TC = Total cholesterol

[*] True predicted exclusions

[†] False predicted exclusions

[‡] Difference in costs per randomization compared with the initial selection strategy

[#] Difference * 8,657.

randomisations (343 of the first 2,200 patients were already randomised). Using the threshold of the fitted model of visit two to predict high HDL at visit three could save about US $ 260,000. Using the model of visit one to predict low TC and excluding everyone with a probability of 0.4 or higher would save about US $ 9,000. However, these savings were not significantly different from zero. Application of the reduced visit two model would save over US $ 147,000. Application of reduced model one and two to predict low TC would save US $ 9,000 and US $ 78,000, respectively. As some participants were excluded by more than one outcome, the total saved costs could not be derived by simply adding the saved costs of the above five selected models. However, combined application of the five models suggested that over US $ 450,000 could be saved over the remaining ROCARI selection period.

Discussion

We have proposed an approach to increase the efficiency of the selection period of a clinical trial by predicting exclusion at subsequent visits using data obtained at earlier visits. If five prediction models were simultaneously applied to the selection period of ROCARI, the costs per randomisation would decrease by US $ 52 and the total screening costs by at least US $ 450,000. Models to predict subsequent exclusions in a selection period can be derived from a pilot study or from interim analyses during patient recruitment.

Large trials like ROCARI[12] or the recently started Women's Health Initiative trial[17] as well as smaller trials with multiple expensive or invasive measurements to determine eligibility[18] tend to apply a selection period. Such trials are, therefore, likely to benefit from the above approach. In the last decade primary prevention trials have received much attention in the literature with a particular emphasis on the reduction of risk for cardiovascular disease. Consequently, the required size of such trials and the number to be screened tend to be very large. The associated high costs could affect their feasibility.[5,19] Application of the proposed method could improve the financial feasibility of primary prevention trials. However, to achieve maximum benefit, the approach should be applied as early as possible in the selection period, preferably based on data obtained from a pilot study. Prediction models derived from a pilot study can be applied to the actual selection period only if the measurements of determinants and outcome considered in the models have not been changed or moved to another visit. Clearly, the more data acquired during the selection period the better the model can be refined and adjusted. This will further enhance their precision and benefits. Derivation or adjustment of the prediction models from interim analysis whilst the selection period is in progress may result in a shift of the distribution of the characteristics at baseline (randomisation). The extent to which this occurs only depends on the falsely predicted exclusion of subjects that would indeed be randomised if the prediction model were not applied. Since the comparison remains based on random allocation this will not affect the internal validity. We also believe that this will not affect the generalisability of the trial results because generalisability will be determined by the eventual distribution of the baseline characteristics. A shift in the distributions at baseline, due to interim application of the prediction models, did not occur in the ROCARI example (table 3.3.4). This is due to the relatively few false predictions.

Table 3.3.4 Characteristics at randomization in the initial situation and after application of the five prediction models (table 5.3), based on data of the cohort of 6,544 men that could have completed the selection period.

	Initial situation without application of the prediction models (N=985*)	New situation with application of the prediction models (N=948*)
Characteristic	Mean	Mean
Age (years)	52.3 (7.4)†	52.4 (7.4)
Mean total cholesterol level (mmol/l)	6.9 (0.6)	6.9 (0.6)
Mean HDL level (mmol/l)	1.1 (0.2)	1.1 (0.2)
Triglyceride level (mmol/l)	2.0 (0.7)	2.0 (0.7)
LDL level (mmol/l)	4.9 (0.6)	4.9 (0.6)
Diastolic blood pressure (mm Hg)	86 (11)	86 (11)
Systolic blood pressure (mm Hg)	135 (18)	135 (18)
Body mass index (kg/m^2)	26.5 (2.9)	26.5 (2.9)

* Number of patients randomized
† Number between parentheses is the standard deviation

In ROCARI, removal of all variables except one from the initial model at visit two resulted in the same savings compared to the initial model. The multivariable models did not, therefore, yield additional information to (univariable) application of a single determinant-threshold. This may be different in other trials.

In the prediction of subsequent exclusions, the true positive predictions directly decrease the costs per randomisation. The false positive predictions are a loss of potential randomisations. As this loss means that an extra number has to be screened in order to obtain the required number of patients in the trial, the false predictions indirectly increase the costs per randomisation. However, the decrease in costs per randomisation over probability thresholds of a prediction model follows an optimum curve (Figure 3.3.2). Therefore, thresholds with either a small number of false positives or a large number of true positives do not correspond to a major decrease in costs per randomisation per se. Furthermore, an upper limit for the number of false positive predictions may exist. This depends on the trial. In trials for which participants are hard to find because the source population is limited, loss of potential randomisations becomes more serious. In such situations, one may prefer a smaller number of false positive exclusions above a higher efficiency. If patients are not hard to find probability thresholds with the highest efficiency are favoured. In our example, the source population which included all male inhabitants of Rotterdam and surrounding communities, was large enough.

Besides the *number* of true and false positives which is a result of the strength of the association between the prediction model and the exclusion to be predicted, several other factors also determine the impact of the model on the efficiency of the selection period. These are briefly discussed. First, the chronological position of the visit from which the model is obtained as well as the costs of the subsequent visits determine the extent to which true predictions decrease the costs per randomisation. The earlier the visit at which the prediction model is obtained and the higher the costs of its following visits, the more measurement costs will be saved. Second, the chronological place of the visit at which the outcome to be predicted is measured determines to what extent the false positive predictions increase the costs per randomisation. Obviously, if the outcome is measured late in the selection period fewer visits are left before randomisation takes place. Hence, the false positive predictions are more likely to be a loss of randomisations because the probability to become excluded after the outcome becomes low. This may also increase the costs per randomisation.

However, a stepwise exclusion process with relatively expensive measurements and low exclusion probabilities at later visits is commonly chosen in the selection period regardless. The proposed approach should, therefore, focus on the strength of the association between the prediction model and the outcome, while the outcome should be predicted as early as possible.

The efficiency may be further increased if the observed value of the predictors are entered in the derived models immediately at the moment of measurement. The patient could directly be withdrawn. Consequently, the costs of the remaining measurements of that particular visit could be saved and the subject would be spared the remaining measurements as well. Besides prediction of subsequent exclusions, the proposed method may also be used to decide on transfers of eligibility criteria with high exclusion rates to earlier visits and criteria with low exclusion rates to later visits, or an adjustment of eligibility criteria. This may further increase efficiency. However, interim adjustment of eligibility criteria may affect the homogeneity of the study population. Depending on the extent to which the adjustment causes a shift in the distribution of the baseline characteristics, this may have consequences for the generalisability. Nevertheless, the generalisability will anyway be determined by the eventual distribution of baseline characteristics.

In conclusion, data obtained early in a selection period of a clinical trial may be used to predict subsequent exclusions. This may increase efficiency of the patient recruitment. All trials which make use of a selection period to recruit the eligible participants can benefit from the proposed strategy. In view of the limitations and recommendations discussed, additional research could refine this method to further enhance efficiency.

References

1. Knipschild P, Leffers P, Feinstein AR. The qualification period. J Clin Epidemiol 1991;44:461-4.

2. The West of Scotland Coronary Prevention Study Group. A coronary primary prevention study of Scottish men aged 45-64 years: trial design. J Clin Epidemiol 1992;45:849-60.

3. Neaton JD, Grimm Jr RH, Cutler JA. Recruitment of participants for the Multiple Risk Factor Intervention Trial (MRFIT). Contr Clin Trials 1987;8:41S-53S.

4. The Scandinavian Simvastatin Survival Study Group. Design and baseline results of the scandinavian simvastatin survival study of patients with stable angina and/or previous myocardial infarction. Am J Cardiol 1993;71:393-400.

5. Silagy CA, Campion K, McNeil JJ et al. Comparison of recruitment strategies for a large-scale clinical trial in the elderly. J Clin Epidemiol 1991;44:1105-14.

6. Hansson L, Dahlof B, Ekbom T et al. Key learnings from the STOP-Hypertension Study: an update in the progress of the ongoing Swedish study of antihypertensive treatment in the elderly. Cardiovasc Drugs Ther 1991;4 (suppl 6);1253-5.

7. MRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis. Lancet 1991;337 (8752):1235-43.

8. Gardner MJ, Heady JA. Some effects of within person variability in epidemiological studies. J Chron Dis 1973;26:781-99.

9. Rotterdam EP, Katan MB, Knuiman JT. Importance of time interval between repeated measurements of total or high-density lipoprotein cholesterol when estimating an individual's baseline concentrations. Clin Chem 1987;33:1913-5.

10. Natelson BH, Tapp WN, Munsif A, Burns W. Fluctuating serum cholesterol: implications for coronary prevention. Lancet 1988;2:404-5.

11. Thompson SG, Pocock SJ. The variability of serum cholesterol measurements: implications for screening and monitoring. J Clin Epidemiol 1990;43:783-9.

12. Rotterdam Cardiovascular Risk Intervention Study (ROCARI) protocol. Rotterdam: Rotterdam Medical Research Foundation; 1990.

13. Rao CR. Linear statistical inference and its applications. New York: Wiley & Sons; 1973:388-9.

14. Harrell FE. The LOGIST procedure. In: Hastings RP, ed. SUGI supplemental library user's guide. Version 5 edition. Cary, NC: SAS Institute Inc;1986:269-93.

15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

16. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.

17. Women's Health Initiative (WHI) protocol. Seattle, Washington: WHI Clinical Coordinating Center, Fred Hutchinson Cancer Research Center; September 1994.

18. Fox K, Pool J, Vos J, Lubsen J. The effects of nisoldipine on the total ischemic burden: the results of the ROCKET study. Eur Heart J 1991;12:1283-7.

19. Pocock SJ, Thompson SG. Primary prevention trials in cardiovascular disease. J Epidemiol Comm Health 1990;44:3-6.

CHAPTER 4

GENERAL DISCUSSION

A central theme in this thesis is the need to take the context of clinical practice into account in diagnostic research or test evaluation. When disregarded this may lead to erroneous conclusions about the clinical relevance of a test. Limitations of prevailing concepts in diagnostic studies have been evaluated and various methods for improvement suggested. The aim of the thesis is to narrow the gap between diagnostic problems in clinical practice and results from diagnostic studies.

In this final chapter the major differences between diagnostic practice and diagnostic research (studies), and inferences from our findings are discussed. For this purpose, we will first review principles of diagnostic practice and research. Secondly, implications of the proposed principles of diagnostic research are discussed from a clinical perspective. Finally, remaining (theoretical) issues that require attention in future diagnostic research are mentioned.

Principles of diagnosis in clinical practice

In the practice of medicine, diagnosis is not an aim in itself. It is relevant in so far as it directs treatment and helps to predict prognosis of a patient presenting with a particular problem.[1] Although there may be a number of possible underlying states of health which could have caused the patient's problem, in the diagnostic work-up a physician commonly addresses a particular disease, i.e. the target disease.[2,3] Which of the potential diseases is defined as the target disease is usually determined by the severity of that disease or by the probability that it may be present. A diagnosis often reflects a dichotomy in which the aim is to assess the probability of the presence or absence of the target disease, in order to initiate treatment or not. To this aim, a physician describes the patient profile by obtaining diagnostic information from the patient. As much diagnostic information is obtained as needed to consider the target disease present or absent with a sufficient degree of confidence. The fundamental purpose of diagnostic testing is to reduce uncertainty about the presence or absence of the disease in order to reduce the risks of an improper treatment decision. The diagnostic information is obtained in a stepwise manner and after each step intuitively expressed into a probability of the presence of disease. A physician "estimates" a probability of the presence of disease using information from the patient history and physical examination and, if necessary, adds information from other diagnostic tests. To set a diagnosis is a multivariable concern per se in which the diagnostic probability is consecutively updated until a therapeutic decision can be made or otherwise sufficient prognostic knowledge is available. This decision is also determined by the costs (or risks) of the diagnostic procedures itself and by the costs of possible misclassification of disease status. In clinical practice, estimation of diagnostic probabilities and specification of the costs and benefits of subsequent therapeutic actions are implicitly integrated into a complex decision process.[3]

In summary:

1. To set a diagnosis is a phased, multivariable process in which the probability of the presence of disease is continuously updated when new diagnostic information is added to the patient profile. This process is continued until a treatment decision can be made.
2. The decision on patient treatment (including "no" treatment) is determined by the diagnostic probability and the costs and benefits of subsequent clinical action.

Principles of diagnostic research

Diagnostic research is inherently descriptive: its only motive is to improve prediction of the presence or absence of disease in order to guide therapeutic decisions. The principles of a diagnostic study may consequently be summarized according to well known concepts in epidemiologic research.

Aim The aim of diagnostic research is to evaluate what diagnostic information is relevant and contributes to the estimation of diagnostic probabilities to the extent that they direct treatment decisions. In other words, to estimate diagnostic probabilities with sufficient precision as necessary, and with minimum patient burden and measurement costs. The amount of necessary diagnostic information on the one hand and the patient burden and measurement costs on the other hand must be balanced. Considerations of efficiency motivate diagnostic research. The objective is to define the most efficient diagnostic function. A diagnostic function describes the prevalence of a particular disease as a joint function of its (diagnostic) determinants.

Study population The patients are selected on the presence of a certain problem or indication. The indication to set a diagnosis defines the clinical domain of a study as well as the target disease.

Determinants and treatment Given the clinical domain and the target disease, potential diagnostic determinants and possible treatment (with their corresponding risks and benefits) can be defined. All potential diagnostic determinants are documented. The presence of disease must be independently, without knowledge of other diagnostic information, determined by a reference standard.

Analysis In clinical diagnosis, information is usually obtained in a sequence of steps. In various phases, for example in patient history, physical examination, and blood and imaging testing, numerous diagnostic variables are documented. It is of clinical relevance to know whether the information on a certain variable, e.g. the result of a particular test, contributes to the diagnosis given the information that is obtained in previous phases: the added value of the diagnostic variable is at interest. For two variables that are documented at the same point in the clinical work-up, the difference in added diagnostic information is of relevance or, in case of equal informativeness, the difference in costs or burden to the patient. Diagnostic studies should evaluate each variable or test within its phase in the clinical work-up. The chronological order by which the value of diagnostic variables are documented determines their hierarchy in the analysis. To evaluate the added value of a particular variable, adjust for complex mutual dependencies with other variables and validly estimate diagnostic probabilities, data must be analysed by multivariable logistic regression modelling.[4,5] Per phase, the independent contribution of (the information on) a variable to the prediction of the presence of disease can be evaluated. Per phase, the most efficient diagnostic model or function can be constructed. If it is possible to define the risks and benefits of subsequent therapeutic decisions[6-8], or the ratio of the net risk of missing a diseased patient to the net risk of treating a non-diseased patient[9,10], the independent contribution of (the information on) each variable to patient management and prognosis can also be evaluated.[11]

Application in diagnostic research

Diagnostic studies are undertaken for a number of reasons; to evaluate the value of current practice in a particular diagnostic problem, to obtain an optimal use of the existing diagnostic arsenal by eliminating (redundant, expensive or invasive) tests, to add a new test to the available diagnostic arsenal, to use an existing test for a new clinical indication, or to replace an existing test by a newer one.[3] Also, when a new treatment with its corresponding risks and benefits becomes available, it may be necessary to study again the diagnostic value of existing tests. For diagnostic testing to have clinical relevance it should contribute to treatment decisions[11]; new treatments may alter the value of particular diagnostic tests. We believe that application of the proposed research principles in any of the above diagnostic studies promotes the clinical relevance of their results. For some of the purposes mentioned this has been illustrated in the thesis. Chapter 2.2 and 3.1 describe studies which address diagnosis in patients with a particular problem in order to select the relevant determinants of the presence of disease from a scala of routinely documented diagnostic variables. Although each diagnostic variable or test may provide information it is likely that, in certain patient groups, some diagnostic tests are partly or completely redundant to previous information, as for example obtained from patient history and physical examination. Moreover, as diagnostic tests may be burdening for the patient, time consuming, expensive and even may produce adverse effects, it is important to restrict the diagnostic process to the relevant procedures only. Similarly, in studies which investigate the value of a new test for a particular clinical problem (chapter 3.2) the added diagnostic information is at issue, or, in case of equal informativeness with existing tests, the difference in costs or burden to the patient.

The principles of diagnostic research can be applied to all clinical settings in which the aim is to predict the presence or absence of a particular state (of health). In chapter 3.3 we have shown that prediction of patient characteristics which determine the inclusion or exclusion in a clinical trial using previously obtained data, may substantially improve the efficiency of the patient selection for clinical trials. In this context, the inclusion and exclusion criteria are used to set the diagnosis "eligible for the trial".

Prevailing diagnostic research

The large majority of studies on diagnostic tests has followed an approach in which the value of a single test to discriminate between the presence and absence of a particular disease is evaluated without reference to its clinical context or to therapeutic consequences. Study results are usually expressed by sensitivity, specificity, or the area under the ROC-curve (ROC area) and should be interpreted with caution. This is only partly because the test parameters are conditional on the presence or absence of disease, whereas diagnostic practice starts from the presence of symptoms and signs. After all, starting from "reverse" probabilities as sensitivity and specificity, Bayes' theorem may be used to estimate the direct diagnostic probability.[2,9] Single test parameters may be deceptive because of their variation according to various clinical and non-clinical patient characteristics. As illustrated in chapter 2.1, this is because results obtained from diagnostic tests and many other patient characteristics are mutually dependent and provide to some extent overlapping information. In fact, rather than "the" sensitivity and specificity of a test for patient subgroups different sensitivities, specificities and ROC areas of a particular test may apply. However, these subgroup specific test parameters are no objects of general medical knowledge.[5] When attempting to estimate a diagnostic probability in an individual patient it is unclear which value of test parameters should be used in Bayes' theorem. Therefore, the published sensitivity, specificity and ROC area of a test should be interpreted with caution as they may not directly speak on the clinical relevance of the test (chapter 2.2).

Missing values

In studies on the evaluation of diagnostic tests, data on test results or on the presence or absence of disease may be missing or uninterpretable. Excluding such data from the analyses may bias the estimates of sensitivity and specificity.[12,13] However, if the missing or uninterpretable test results are equally distributed among the diseased and non-diseased patients and if it is unlikely that the cause of such results is related to the potentially observable test result, it will probably not affect (inferences on) the diagnostic value of the test. Similarly, if the missing data on the disease status are equally distributed across the spectrum of test results and if it is unlikely that the reason for missing the data is related to the true disease status, it will not affect the study results. In diagnostic studies, information on missing data is often lacking but its potential impact on the study results can and should be discussed.

Screening

Research on screening tests regards a particular case of diagnostic studies. Typically, a screening test provides information to assess the probability of presence of a particular disease in its early, developing stage. Other patient information is neglected except perhaps for a particular age range and gender. A chronological hierarchy of variables and the added value of the test is not at issue. Therefore, the test result distributions among diseased and non-diseased patients or parameters like the ROC area, sensitivity and specificity do not have the limitations they have for clinical diagnosis and may be used to indicate the test's screening value or to compare screening tests. If a subject is referred for further *diagnostic* work-up because of a positive screening test result, however, aspects of conditionality start to play a role. Consequently, the independent or added value of diagnostic tests become important.

Generalisability and validation

This thesis aims to narrow the gap between diagnosis in clinical practice and the design and methods of analysis in diagnostic studies. However, we have concentrated on the need to consider the context of clinical practice in diagnostic studies rather than on the application of results from diagnostic studies in clinical practice. With the latter we refer to the implementation of a diagnostic function, often simplified to a prediction rule, into practice.[14-16] Such implementation requires adequate validation of the derived diagnostic function. A particular diagnostic function can discriminate well between the presence and absence of the disease in the study population but may be unreliable elsewhere.[17,18] When constructing a diagnostic function all available variables that modify the disease probability estimation and satisfy other efficiency considerations regarding measurement costs and patient burden are included. Population differences in these variables will not affect generalisability and will not reduce performance of the function. Problems, however, may arise due to population differences in variables that are unknown or removed from the function during its derivation and interact with the variables included. Therefore, before application in clinical practice, diagnostic functions should be tested in other patient populations selected from the same clinical domain as for which the function was derived. The split sample method, as we have applied in chapter 3.3, is an alternative, less demanding method to validate a prediction model.

Another problem that may arise with the application of a (validated) diagnostic function in practice is that a particular diagnostic test can not be performed or is not available. One way to handle this problem is to use the (presumed) mean test result for the corresponding patient subgroup. This mean can be obtained from the literature. However, this kind of data is hardly reported in diagnostic studies. Another way to handle the problem is to apply the diagnostic function minus the particular test. To this aim, diagnostic studies should report all subsequent diagnostic functions according to the chronological phases of the diagnostic work-up. This enables physicians to select the diagnostic function(s) that can be applied to their clinical setting.

The "gold" standard

The extent to which results from diagnostic procedures have an impact on the objective or subjective course of the disease (clinical outcome), such as reduction in morbidity or mortality, decrease in time to recovery, improved quality of life or an improved cost-effectiveness, reflect the eventual clinical value of a diagnostic procedure.[11,19] Ideally, diagnostic studies evaluate the test's ability to influence, albeit indirectly, clinical outcomes. Because such studies often require a much more extensive follow-up, diagnostic tests are usually evaluated using a so-called "gold standard" which is supposed to represent the diagnostic "truth", i.e. the true presence or absence of disease. Since any definition of truth may lead to rather philosophical discussions, it should be appreciated that the standard test does not have to be "24 carat" gold. Any test that is applied in practice to exclusively and ultimately direct patient management, can serve as a gold standard test in diagnostic studies. The operational "truth" of clinical practice offers a sufficient reference standard for diagnostic research. In this context, the term "reference test" would be more appropriate. Evaluation of a diagnostic model or test to a reference test instead of a gold standard will generally yield lower limits of the sensitivity, specificity and predictive value of the model. This is due to misclassification on true disease status by the reference test. When a better reference test comes available the value of these parameters will generally increase. Therefore, in diagnostic problems which lack a ("24 carat") gold standard the value of diagnostic parameters should be regarded as temporary. Also note that theoretically the interpretation of these parameters is different as they do not reflect the true presence and absence of disease anymore.

The theoretical considerations of diagnostic research as outlined in this thesis apply to clinical problems for which a reference test or gold standard is available. In these instances, judgement on the diagnostic "truth" by the reference test or gold standard must be performed independently from the information provided by the diagnostic variables that are to be studied.[7,20] For clinical problems which lack such operational, independent, reference test or gold standard, the truth can better be expressed in terms of clinical outcome. Although this may be more difficult to achieve, in terms of time and efforts, it may yield more relevant and valid information for practice.[3]

Randomised trials in diagnostic research

To empirically evaluate the impact of the results of any (new or investigational) diagnostic procedure on clinical outcomes, randomisation is required to prevent confounding by indication.[20-23] There are several randomised trial designs possible in which the moment of randomisation may vary. Probably the most efficient design is to independently perform both the conventional diagnostic procedure and the new diagnostic test on each patient. Two diagnostic conclusions are obtained: one conclusion based on information without and one based on information with the new test. Only the patients with different diagnostic conclusions that would lead to a different treatment strategy (including no treatment) need to be randomised. Irrespective of the diagnostic information obtained, randomisation takes place between either the treatment choice according to the conventional procedure or the treatment choice according to the new procedure. At the end of the trial we may define specific patient subgroups according to their diagnostic information that benefit or not from a particular treatment. Additionally, this design allows for detection of differences in treatment effects according to similar diagnostic profiles. However, this design is only ethical if the combined information of the conventional and new procedures does not result in a much better classification of disease presence as compared to the individual tests, and if both treatment strategies allow for randomisation.

If it is not ethical to randomise between the two treatment strategies, a more pragmatic approach would be to randomise the patients to one of the two (or more) combined diagnostic and therapeutic strategies. In the reference group patients are subjected to the conventional diagnostic procedure and the subsequent treatment as directed by the obtained information. In the study group patients are subjected to the new or investigational diagnostic procedure and the subsequent treatment as directed by the obtained information. Note that this treatment can be the same as in the reference group. In both arms, the diagnostic procedures and therapy are evaluated in combination. Although with this design the effect of the test may not be distinguished from the treatment effect, an improved prognosis of the disease in the intervention arm gives evidence that the new procedure provides better guidance for treatment. However, this design requires prior knowledge about the therapeutic information provided by the new test, i.e. the test results that indicate or contra-indicate the treatment.

In both designs the investigational test should be carried out in addition to diagnostic variables that are employed regardless, e.g. patient history and physical examination. Both designs allow for a valid comparison of the investigational test to a particular test, to a combination of tests, and to evaluate the added value of the new test to the conventional test. However, in the pragmatic approach one has to choose beforehand two (or more) particular diagnostic strategies to be compared whereas the former, more explanatory approach, allows for evaluating all possible combinations of diagnostic variables, i.e. a diagnostic function, in order to select the most sufficient one. Another advantage of the first design is that it gains in efficiency due to the paired observations and the exclusion of patients with similar diagnostic conclusions that would initiate the same therapy. However, if there is little prior knowledge about the therapeutic information provided by the new test, *all* patients may still need to be randomised between possible treatment strategies.

A major advantage of the randomised trial in diagnostic research is that the ability of a diagnostic procedure to correctly predict the presence of the disease (given the availability of a reference test), to direct patient management as well as to modulate clinical outcomes, can be evaluated validly and simultaneously. However, randomised trials are expensive and can logistically be complex. Therefore, they have rarely been conducted in diagnostic research. Emphasis remains on studies on prediction of the disease presence.[19,21,22] Accordingly, this thesis has focused on the principles of diagnostic research in predicting a patient's health status. It should be appreciated that if an indicator does not contribute to this prediction it is unlikely that it will affect treatment decisions or clinical outcome. Similarly, if the (added) diagnostic information of a new test with respect to patient burden and costs is not superior to existing procedures, a trial to evaluate clinical outcome is irrelevant. Alternatively, a test may improve prediction of the presence of disease without affecting subsequent therapeutic consequences. In chapter 4, we explored an approach which allows to obtain information on the test's ability to improve prediction of disease presence and to affect treatment decisions without the need for a randomised trial. This approach evaluates diagnostic tests with application of the treatment probability thresholds as defined by the costs and benefits of the treatment.[6-8] If on average the test results increase or decrease the prior probability of disease presence above or below this threshold, the test may affect therapeutic actions. If previous trials have proven that the treatment at issue is efficacious, or that its withholding prevents harmful effects in certain patients, the beneficial effect of the test on clinical outcome may be considered as established and, in fact, be estimated.[21]

Future diagnostic research

In this thesis principles of diagnostic research have been discussed. It was demonstrated that application of these principles may improve the clinical relevance of diagnostic study results. Accordingly, future studies on diagnostic probability estimation should apply these concepts. We have also explored an approach to evaluate a test's ability to affect treatment decisions. This provides an indirect method to obtain knowledge about the test's contribution to improve clinical outcome, which reflects its true value for medical practice. The current view is that the latter can only validly be studied in a randomised study. It is of interest to further develop study methods that provide knowledge on the true clinical value of a test without the need of a randomised trial. Theoretical improvements as attempted in this thesis for studies on prediction of disease presence, should receive particular attention in future diagnostic research.

Commonly in diagnostic research estimation of the probability of the presence of a particular disease, i.e. the target disease, forms the objective. However, there may be several underlying "diseases" that potentially caused the problem and these differential diagnoses will always be considered by a physician. This thesis has focused on diagnostic models which describe the relation of one or more determinants to the presence or absence of the target disease. In these diagnostic models the prevalence of alternative diseases was not considered. The availability of (statistical) methods to simultaneously evaluate the relation of diagnostic determinants to the prevalence of several diagnoses or outcomes would further increase clinical relevance of diagnostic studies. Polychotomous logistic modelling may provide an appropriate tool. It should be realised, however, that such studies require larger study populations as the number of diagnostic categories increases.

This thesis has attempted to provide a framework for diagnostic research. A sufficient methodologic and statistical basis for this kind of clinical epidemiologic research is still lacking. Future studies should improve the theoretical basis of diagnostic research and solve remaining problems such as those raised in the discussion. As diagnosis forms the basis of clinical medicine, would it not be time to know how to evaluate diagnostic tests properly, just as we have come to learn how to validly study efficacy of treatment ?

References

1. Habbema JDF. Clinical decision theory: the threshold concept. Neth J Med 1995;47:302-7.

2. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology; a basic science for clinical medicine. Boston: Little, Brown & Co, 1985.

3. Zweig M, Campbell G. Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(4):561-77.

4. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313:793-9.

5. Miettinen OS, Caro JJ. Foundations of medical diagnosis: what actually are the parameters involved in Bayes' theorem? Stat Med 1994;13:201-9.

6. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229-34.

7. Metz CE. Basic Principles of ROC analysis. Semin Nucl Med 1978;8:283-98.

8. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109-17.

9. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia; W.B. Saunders, 1980.

10. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworths, 1988.

11. Asch DA. Patton JP, Hershey JC. Knowing for the sake of knowing: the value of prognostic information. Med Decis Making 1990;10:47-57.

12. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chron Dis 1986;39:575-84.

13. Simel DL, Feussner JR, Delong ER, Matchar DB. Intermediate, indeterminate and uninterpretable test results. Med Decis Making 1987;7:107-14.

14. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. New Engl J Med 1985;313:793-9.

15. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. Ann Intern Med 1990;113:664-70.

16. Bates DW; Cook EF, Goldman L, Lee TH. Predicting bacteremia in hospitalized patients. A prospectively validated model. Ann Intern Med 1990;113:495-500.

17. Diamond GA. What price perfection? Calibration and prediction of clinical prediction models. J Clin Epidemiol 1992;45:85-9.

18. Harrell Jr FE, 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.

19. Begg CB. Experimental design of medical imaging trials: issues and opinions. Invest Radiol 1989;24:934-6.

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

21. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-94.

22. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071-81.

23. Köbberling J, Trampisch HJ, Windeler J, eds. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990;28:873-9.

CHAPTER 5

SUMMARY

Diagnosis forms the basis for patient care. Diagnosis is not an aim in itself but is relevant in as far as it directs treatment and indicates the prognosis of a patient. Diagnosis amounts to an estimation of the probability of the presence of a disease in view of all diagnostic information (patient history, physical examination and test results) in order to decide whether treatment should be initiated or not. A diagnosis is rarely based on one single variable or test and therefore is a multivariable concern per se. However, most diagnostic research or studies in which diagnostic tests are evaluated still follow a univariable approach. This means that a test is evaluated in isolation without explicit regard to the clinical context in which the test will be applied. In this respect, clinical practice and diagnostic studies frequently do not cohere. The medical literature has paid too little attention to these principles of diagnostic research. This thesis endeavours to outline the principles of diagnostic research in order to improve its clinical relevance. The general aim is to narrow the gap between diagnostic problems in clinical practice and results from diagnostic studies.

Chapter 2 explores limitations of prevailing methods and concepts in diagnostic studies and describes some alternative approaches. In chapter 2.1 the relevance of the conventional diagnostic test parameters such as sensitivity, specificity and likelihood ratio (LR) in clinical diagnosis is evaluated. Using diagnostic data of 295 patients suspected of coronary artery disease (CAD) who all underwent coronary angiography (reference standard) it was shown that these parameters of the exercise test substantially varied according to patient characteristics obtained from patient history, physical examination and measures of disease severity. As each patient population tends to be heterogeneous with respect to patient characteristics, we propose that no single level can be given of these parameters of any diagnostic test that is adequate for all patient subgroups. Therefore, clinical use of these parameters as a basis for calculating diagnostic probabilities in individual patients, even when using Bayes' theorem, has serious limitations. Such probabilities can validly be estimated using multivariable logistic regression models.

Chapter 2.2 illustrates the hazards of a univariable approach in the evaluation of diagnostic tests with respect to their clinical application. In 140 patients suspected of pulmonary embolism who already had an inconclusive (intermediate) ventilation-perfusion lung scan result we evaluated the diagnostic value of leg ultrasound, chest X-ray and arterial oxygen pressure. In univariable analyses the positive predictive value, LR and specificity of leg ultrasound was markedly higher than of arterial oxygen pressure and chest X-ray whereas the latter had a substantially higher sensitivity. However, in multivariable logistic regression models we found that arterial oxygen pressure had no added value to patient history plus physical examination and the added value of chest X-ray and leg ultrasound was similar. As these tests are always performed after findings from patient history and physical examination are available, it is this added value that marks clinically relevant information. Accordingly, it was concluded that diagnostic parameters of a single test may not be indicative for its clinical potential. A 'threshold of diagnostic relevance' for these parameters can hardly be given. Therefore, diagnostic studies should follow a multivariable logistic regression approach and studies which evaluate the diagnostic accuracy of a test in a univariable way should be interpreted with caution.

The performance of diagnostic tests and (multivariable) diagnostic models is often compared using the area under Receiver Operating Characteristic (ROC) curves. In chapter 2.3 we contrast this approach with a more direct method that takes into account therapeutic consequences of a diagnosis. Diagnostic data obtained from the same 140 patients as described in chapter 2.2 were used for an illustrative example. We showed that two diagnostic models with the same ROC area may perform very differently when costs (or risks) and benefits of subsequent decisions are considered. Two other models had substantially different ROC areas but performed similarly taking into account their therapeutic consequences. Comparison of diagnostic tests simply using the ROC areas may lead to erroneous conclusions about therapeutic utility. It would be more appropriate when the test's clinical implications are also considered in diagnostic test evaluation. This is feasible by explicit definition and application of a treatment threshold which is proposed in this study.

Chapter 3 describes three diagnostic studies in which the theoretical considerations of chapter 2 are taken into account. Among 451 patients suspected of pulmonary embolism, chapter 3.1 evaluates the diagnostic value of patient history and the added value of physical examination, arterial blood gas values, chest radiography and perfusion scintigraphy, according to the chronological order in which data comes available in practice. This was done by systematically constructing and extending multivariable diagnostic models. We found that independent predictors obtained from patient history (age, recent surgery, previous deep venous thrombosis, dyspnoea, collapse) and physical examination (pleural rub, signs of deep venous thrombosis, breath frequency) contribute to the confirmation or exclusion of presence of pulmonary embolism. Given this diagnostic information, blood gas values and chest radiography provide limited additional information. When prior knowledge is regarded these tests are redundant to assess the presence of pulmonary embolism in terms of diagnostic efficiency (improvement of disease prediction, burden to the patient and measurement costs). The added value of perfusion lung scanning, however, was substantial.

Chapter 3.2 describes a study on continuous ST-monitoring testing, which measures the extent and duration of myocardial ischaemia, to predict the enzymatic infarct size and left ventricular function in patients with acute myocardial infarction. This was studied among 269 patients who received thrombolytic therapy and were enrolled in the ECG monitoring substudy of GUSTO-I. Using linear regression models, we showed that the area under the ST-trend until 50% ST-recovery and the total area under the recurrent ischemic episodes, are predictors of the enzymatic infarct size and left ventricular function, independent from other patient characteristics, e.g. age and infarct location. This supports the clinical application of ST-segment monitoring tests to provide (diagnostic) information about the cardiac condition of patients with acute myocardial infarction. ST-monitoring tests may, therefore, guide the physician to tailor thrombolytic therapy. Further research is necessary to evaluate their true (independent) value to physicians and to show whether they truly lead to better therapeutic interventions and prognosis.

To determine eligibility for a (randomised) clinical trial, measuring the inclusion and exclusion criteria can be extended over a period of time. During this period, known as the selection period, a patient is repeatedly examined at certain time intervals before the patient is randomised at the end of the period. In fact, the selection period is performed in order to set the diagnosis "eligible for the trial". Chapter 3.3 describes a study in which we applied the research principles as outlined in previous chapters to increase the efficiency of the selection period of a large primary prevention trial on the efficacy of a cholesterol lowering drug. We showed that data obtained at early examinations in the selection period could predict subsequent exclusions. In the example, application of only a few multivariable prediction models to the selection period would already decrease the total costs per randomisation by US $ 450,000 compared to the situation without using the models. Data obtained in a pilot study as well as data obtained in the beginning of a prolonged selection period may be used to construct predictive algorithms in large randomised trials or in trials with extensive, invasive or expensive selection measurements.

In chapter 4 the implications of the proposed principles are discussed from a clinical perspective. The thesis has attempted to construct a framework of methods for studies to evaluate the contribution of diagnostic determinants to the estimation of disease probabilities. It was demonstrated that application of these methods may improve clinical relevance. Accordingly, in future studies these concepts should be considered. We have also explored an approach to evaluate a test's ability to affect treatment decisions. This provides an indirect method to obtain knowledge about the test's contribution to improve clinical outcome, which reflects its true value for medical practice. The current view is that the latter can only validly be studied in a randomised study. It is important to further develop study methods that provide knowledge on the true clinical value of a test without the need of randomised trials. Theoretical improvements as attempted in this thesis for studies on prediction of disease presence, should receive particular attention in future diagnostic research.

CHAPTER 6

SAMENVATTING

Diagnostiek is geen doel op zichzelf maar verschaft belangrijke informatie over de therapeutische (contra-)indicaties en prognose van de patiënt. Het stellen van een diagnose is het schatten van de kans op aanwezigheid van een bepaalde ziekte op basis van alle beschikbare diagnostische informatie (anamnese, lichamelijk onderzoek en testresultaten) teneinde een therapeutische beslissing te kunnen nemen. Een klinische diagnose is zelden gebaseerd op één enkele diagnostische indicator of test. Elke indicator maakt deel uit van een scala van indicatoren; diagnostiek is per definitie multifactorieel. Echter, onderzoekingen waarin een diagnostische test wordt geëvalueerd hebben tot op heden vaak een univariabele benadering gevolgd. Dit betekent dat de test afzonderlijk wordt onderzocht zonder expliciet rekening te houden met de klinische context waarin die uiteindelijk wordt toegepast. Op dit punt verschillen de praktijk van de diagnostiek en het wetenschappelijk onderzoek dus aanmerkelijk. Dit proefschrift bespreekt onderzoeksmethoden die de klinische relevantie van diagnostische studies kunnen vergroten en wellicht de discrepantie tussen de praktische diagnostiek en de resultaten van diagnostische studies verkleinen.

Hoofdstuk 2 richt zich op de theoretische en methodologische aspecten van diagnostische studies. Hoofdstuk 2.1 evalueert de klinische relevantie van de conventionele diagnostische testparameters zoals sensitiviteit, specificiteit en likelihood ratio (LR). Dit is geïllustreerd met gebruikmaking van diagnostische gegevens van 295 patiënten met mogelijk coronair vaatlijden. De waarde van genoemde parameters van de fietsproef bleken sterk afhankelijk te zijn van andere patiëntkarakteristieken zoals geslacht, cholesterol gehalte en ernst van de ziekte. Omdat vrijwel iedere patiëntenpopulatie heterogeen is in dergelijke karakteristieken werd geconcludeerd dat één bepaalde waarde voor deze diagnostische parameters voor geen enkele test bestaat. Voor iedere patiëntensubgroep geldt mogelijk een andere waarde. De klinische toepassing van de testparameters voor een valide schatting van individuele diagnostische kansen met behulp van de regel van Bayes is daarom beperkt. Dergelijke kansschatting is wel mogelijk met behulp van multivariabele logistische regressiemodellen.

Hoofdstuk 2.2 beschrijft de gevaren van een univariabele benadering in de evaluatie van diagnostische tests met betrekking tot hun klinische relevantie. Bij 140 patiënten met een mogelijke longembolie en een twijfelachtig resultaat op de ventilatie-perfusie longscan, vergeleken we de diagnostische waarde van drie verschillende tests: echografie van de benen, arteriële zuurstofdruk en de thoraxfoto. Uit univariabele analyses bleek dat zowel de predictieve waarde en LR van de positieve testuitslag als de specificiteit van de been-echografie veel hoger was dan van de andere twee tests, terwijl de thoraxfoto een veel hogere sensitiviteit en lagere LR van de negatieve testuitslag had. Multivariabele logistische regressiemodellen lieten echter zien dat de arteriële zuurstofdruk geen additionele informatie aan de anamnese en het lichamelijk onderzoek gaf terwijl de additionele informatie van de thoraxfoto en been-echografie gelijk was. Aangezien alle drie de testen in praktijk worden toegepast na anamnese en lichamelijk onderzoek is juist deze toegevoegde waarde van klinisch belang. De diagnostische parameters, ofwel de diagnostische informatie, van een test afzonderlijk zegt niet zonder meer iets over de klinische relevantie van de test. Diagnostische studies dienen bij voorkeur een multivariabele benadering te volgen en studies waarin testen afzonderlijk zijn onderzocht dienen met voorzichtigheid geïnterpreteerd te worden.

Hoofdstuk 2.3 vergelijkt de methode van Receiver Operating Characteristic (ROC) curves voor de evaluatie van diagnostische testen en (mulitvariabele) diagnostische modellen met een alternatieve methode die rekening houdt met de therapeutische consequenties van een diagnose. Dit wordt geïllustreerd aan de hand van diagnostische data van dezelfde 140 patiënten als beschreven in hoofdstuk 2.2. De diagnostische betekenis van twee multivariabele modellen welke eenzelfde oppervlakte onder de ROC curve hadden, leek te verschillen wanneer de kosten en baten van de therapeutische beslissing in overweging werd genomen. Twee andere diagnostische modellen hadden daarentegen zeer verschillende ROC curves maar waren van vergelijkbare diagnostische betekenis op basis van hun therapeutische consequenties. Vergelijking van diagnostische testen op basis van ROC curves alleen kan leiden tot verkeerde conclusies over klinische relevantie. Het is van groot belang om de therapeutische implicaties van de testen mee te nemen in diagnostisch onderzoek. Dit is, zoals aangetoond in deze studie, mogelijk middels expliciete definitie en toepassing van de therapeutische beslisdrempel.

Hoofdstuk 3 beschrijft drie onderzoekingen waarin de theoretische concepten uit hoofdstuk twee zijn toegepast. In hoofdstuk 3.1 is bij 451 patiënten met een mogelijke longembolie de diagnostische waarde van anamnese en de toegevoegde waarde van respectievelijk lichamelijk onderzoek, arteriële zuurstofdruk, thoraxfoto en de perfusie longscan onderzocht. In chronologie overeenstemmend met de klinische praktijk, werden hiertoe systematisch multivariabele diagnostische modellen ontwikkeld.

Verscheidene anamnestische variabelen (o.a. leeftijd, recente operatie en kortademigheid) en gegevens van het lichamelijk onderzoek (o.a. pleura wrijven, tekenen van diep veneuze trombose en ademhalingsfrequentie) voorspelden redelijkerwijs de aan- of afwezigheid van een longembolie. De toegevoegde diagnostische waarde van de arteriële zuurstofdruk en thoraxfoto was zeer beperkt. In het kader van diagnostische efficiency (zo goed mogelijk de ziekte voorspellen met acceptabele kosten en lasten voor de patiënt) lijken voor dit doel beide testen overbodig. De toegevoegde diagnostische waarde van de perfusie longscan was wel substantieel.

Hoofdstuk 3.2 beschrijft een studie naar de relatie tussen continue ST-monitoringskarakteristieken, welke een maat zijn voor de mate en duur van coronaire ischaemie, en de enzymatische infarct grootte en linker ventrikelfunctie bij patiënten met een acuut myocard infarct. Dit werd onderzocht bij 269 patiënten die behandeld werden met trombolytica en deelnamen aan de ECG-substudie van de GUSTO-I trial. Met behulp van lineaire regressiemodellen werd aangetoond dat zowel het oppervlak onder de ST-trend tot aan het moment van 50% ST-herstel als het totale oppervlak onder recidiverende ischaemische episodes, onafhankelijke predictoren van de infarct grootte en de ejectiefractie waren. Dit verband bleef bestaan na correctie voor andere patiëntkarakteristieken zoals leeftijd en infarct locatie. Deze resultaten ondersteunen de klinische toepassing van continue ST-monitoringstesten voor het verschaffen van (diagnostische) informatie over de cardiale toestand van patiënten met een acuut myocard infarct. Dientengevolge is ST-monitoring wellicht van waarde voor het coördineren van de trombolyse.

De in- en exclusie criteria voor een clinical trial worden in plaats van gelijktijdig op één patiëntbezoek ook vaak verspreid gemeten over een bepaalde tijdsperiode. Tijdens deze zogenoemde "selectieperiode" worden de patiënten herhaaldelijk gemeten op achtereenvolgende bezoeken om hun geschiktheid voor de trial te bepalen. In een selectieperiode wordt in feite de diagnose "geschikt voor het onderzoek" gesteld. Hoofdstuk 3.3 laat zien dat de bovenbeschreven theoretische aspecten van diagnostische studies tevens gebruikt kunnen worden om de efficiency van een selectieperiode te optimaliseren. Data van de selectieperiode (5 patiëntbezoeken), van een grootschalige, primaire preventie trial naar de effectiviteit van een cholesterolverlagend middel zijn gebruikt ter illustratie. We hebben laten zien dat met behulp van gegevens op de eerste twee patiëntbezoeken bepaalde exclusies op latere bezoeken voorspeld konden worden. Toepassing van vijf ontwikkelde predictiemodellen op deze selectieperiode zou in totaal gemiddeld US $ 450,000 kunnen besparen. We concludeerden dat data verkregen uit een pilotstudie ofwel initiële data uit een voortdurende selectieperiode gebruikt kunnen worden om efficiency verhogende predictiemodellen te construeren. Dit geldt zowel voor grootschalige trials (zoals in het voorbeeld) als voor kleinere trials waarbij erg invasieve of dure procedures in de selectieperiode worden toegepast.

Hoofdstuk 4 beschrijft zowel de klinische implicaties van de onderzochte methoden alsmede hun algemene toepassingsmogelijkheden in diagnostische studies. Vervolgens wordt speciaal aandacht besteed aan de generaliseerbaarheid van studieresultaten, de "gouden standaard" en de gerandomiseerde trial in diagnostisch onderzoek. Tenslotte worden suggesties gedaan voor vervolgonderzoek met betrekking tot de diagnostische theorie en praktijk.

Dankwoord

Nu u zojuist het voorgaande heeft gelezen bent u vast niet verbaast wanneer ik u vertel dat dit proefschrift tot stand is gekomen dankzij de bijdrage van een groot aantal mensen. Het moment is daar om een aantal nog eens extra te bedanken.

Allereerst wil ik mijn promotor professor Rick Grobbee bedanken. Zijn intrinsieke diagnostische test waarmee hij zowel inhoudelijke als taalkundige dwalingen mijnerzijds wist te detecteren, was onmiskenbaar. Of we nu daarboven in de ivoren toren, op een zonovergoten terras onder het genot van een biertje en sigaar, of op een of ander goed verzorgde barbecue vertoefden, hij bleek telkens weer opnieuw in staat licht te werpen op de soms duistere wegen der diagnostiek.

Mijn tweede promotor, professor Dik Habbema heeft mij geleerd hoe ingewikkelde wiskundige formules behalve op vier A4-tjes eveneens op de achterkant van een bierviltje geschreven konden worden. Hij heeft mij ook doen inzien dat de diagnostiek en de klinische besliskunde onlosmakelijk met elkaar verbonden zijn. Alleen al voor deze twee lessen dank ik hem zeer.

Toen mijn co-promotor Gerrit-Anne van Es reeds in het eerste half jaar van mijn promotieonderzoek vertrok naar een andere werkplek, kreeg ik het even benauwd. Echter, mede dankzij zijn intensieve voetbaltrainingen was zijn conditie toereikend om dikwijls op en neer te rennen ten behoeve van mijn begeleiding. Dit heeft geresulteerd in een zeer prettige en vruchtbare samenwerking. Mijn dank hiervoor.

Theo Stijnen wil ik met name danken voor zijn (wis)kundige begeleiding bij het analyseren van vrijwel iedere dataset. Bij hem kon ik altijd binnenlopen en ongegeneerd mijn onwetendheid op tafel leggen.

Jaap Deckers wil ik vooral bedanken omdat zijn medewerking aan dit proefschrift heeft geleid tot het thema van mijn promotieonderzoek.

Missing data zijn altijd een probleem in epidemiologisch (promotie)onderzoek, maar missing datasets laat alles te wensen over. Onmisbaar noem ik daarom de bijdrage van Harry Büller en Bowine Michel. Dankzij hen kan ik dit jaar nog promoveren.

Professor Maarten Simoons, Peter Klootwijk, Simon Meij, Timo Lenderink en Taco Baardman ben ik zeer erkentelijk voor hun belangrijke bijdrage aan het GUSTO-verhaal.

Misschien enigszins onverwacht, maar de gegevens van de onvoltooide ROtterdam CArdiovascular Risk Intervention (ROCARI) trial bleken voor mijn onderzoekingen echter zeer geschikt. Hiervoor wil ik de 'principal investigators' en vertegenwoordigers van ROCARI destijds, professor Rick Grobbee, professor Bert Hofman, Jan Jonker en Annette Bak, nog eens hartelijk bedanken. Tevens dank ik alle ROCARI-medewerkers die zich destijds gigantisch hebben ingezet bij de verzameling en verwerking van de onderzoeksgegevens.

Aangezien een multivariabele regressie analyse niet zonder meer op een rekenmachientje kan worden uitgevoerd is de computer onontbeerlijk voor de moderne epidemioloog. Ditzelfde geldt voor de mensen die de afgelopen jaren in staat waren om mijn vragen op het gebied van computers en datamanagement op te lossen. Zij allen worden bij dezen hartelijk bedankt. In het bijzonder dank ik René Vermeeren en Marcel Eijgermans voor hun hulp bij INGRES- (en auto-) problemen, Chris de Ruiter en Eric Neeleman voor hun hulp bij het construeren van prachtige ROC-curves, en Michael Koenders voor het "charteren" van steeds maar weer zwaardere computers.

I would like to thank professor Olli Miettinen for guidance and inspiration during my thesis. Through the many meetings we had in various hotels and restaurants of Rotterdam he made a great contribution to my work.

Alle dames en de enige heer, professor Hans Valkenburg, met wie ik de afgelopen jaren mijn werkkamer heb mogen delen wil ik bedanken voor hun luisterende oren en helpende handen bij velerlei problemen. Met name dank ik hierbij Bettina Samrèn, Dorothée Wientjens en Caroline van Rossum die dit maar liefst drie jaar lang hebben volgehouden. Hun heerlijke bakjes koffie en zelfs hun gezeur over mijn getelefoneer zal ik zeker missen. Alle andere medewerk(st)ers van het Instituut Epidemiologie & Biostatistiek en van het Instituut Maatschappelijke Gezondheidszorg die mij op enige wijze hebben geholpen of bijgestaan in de voorbije periode wil ik bedanken. Een speciaal woord van dank gaat uit naar twee mensen: Huib Burger en Eric Neeleman. Met Huib was het altijd goed filosoferen over wetenschappelijke, en soms ook, wanneer de tijd en de wijn het toelieten, over wat minder wetenschappelijke kwesties. Eric heeft mij in de laatste fase geweldig geholpen bij het ophelderen van enkele poppetjes op de omslag.

Arno Moons en Huib Burger wil ik bij voorbaat al bedanken voor hun nodige 'back-up' tijdens de verdediging. Een goede steun in de rug zal waarschijnlijk geen overbodige luxe zijn.

Mijn vrienden ben ik zeer dankbaar voor hun niet aflatende steun en belangstelling gedurende mijn A.I.O.-jaren en vooral in de laatste fase daarvan. Zij stonden altijd voor me klaar en bij hen kon ik telkens weer terecht als ik het even niet meer zag zitten of gewoon even stoom moest afblazen. Een klein aantal heeft hierin, misschien onbewust, een zeer belangrijke rol gespeeld en hen wil ik nogmaals bedanken. In dit verband wil ik ook met name mijn zus Lianne, mijn beide broers Michel en Arno, en mijn zwager Ronald (mede voor zijn stellige opmerkingen) bedanken.

Mijn moeder is een geval apart en verdient daarom ook aparte aandacht. Haar wil ik bedanken om het simpele feit dat zij er echt altijd was, wat er ook gebeurde. Mam, bedankt !

Dit boekje was er nooit geweest zonder de steun van Sam. Mijn dank naar haar kan ik niet in woorden uitdrukken. Sam, voor jou heb ik gelukkig ook geen woorden nodig. Het enige dat mij rest te zeggen is: 'm tamob !

Curriculum vitae

Carl Moons was born on March 8, 1967 in Nijmegen, The Netherlands. He graduated in 1984 at the 'Dominicus College' (secondary school) in Nijmegen. He studied 'Gezondheidswetenschappen' (Health Science) at the Catholic University of Nijmegen. As part of his study he wrote a protocol to evaluate the value of ultrasound in diagnosis of tumours and polyps of the colon (department of Epidemiology of the Catholic University Nijmegen, supervision prof.dr. A.L.M. Verbeek) and studied methods to optimise patient selection in clinical trials (department of Epidemiology & Biostatistics of the Erasmus University Rottterdam, supervision prof.dr. D.E. Grobbee). For several years he was a student member of the faculty board of the Catholic University Nijmegen. He graduated in 1991 and started to work for the Rotterdam Medical Research Foundation. In 1992, he began to work on his thesis at the department of Epidemiology & Biostatistics (head: prof.dr. A. Hofman) and the center of Clinical Decision Sciences, department of Public Health (head: prof.dr. J.D.F. Habbema) of the Erasmus University Rottterdam. During this period he received training as an epidemiologist and obtained his MSc-degree in Clinical Epidemiology.

Tabel 2.1.2 **Variations in sensitivity, specificity and LR of the heart rate adjusted ST depression according to various characteristics of patients with and without CAD, expressed in rate difference and likelihood ratio ratio.**

	CAD patients			non-CAD patients				
Patient characteristic	N	Sens* (%)	RD* (95% CI)*	N	Spec* (%)	RD (95% CI)	LR*	LRR* (95% CI)
Patient profile								
Age (years)								
28-50	71	57.8	0.4 (-13.8 to 14.6)	46	91.3	-	6.6	-
51-70	136	57.4	-*	42	92.9	1.6 (- 9.7 to 12.8)	8.1	1.2 (0.3- 5.2)
Sex								
male	170	63.5	33.8 (17.4 to 45.2)†	52	88.5	-	5.5	-
female	37	29.7	-	36	97.2	8.7 (- 1.4 to 19.0)‡	10.7	1.9 (0.2-16.4)
Symptoms								
non-specific	18	50.0	-	37	91.9	3.7 (-14.0 to 21.3)	6.2	1.2 (0.2- 6.9)
atypical	55	50.9	0.9 (-25.7 to 27.5)	34	94.1	5.9 (-11.4 to 23.1)	8.7	1.7 (0.3-11.1)
typical	134	61.2	11.2 (-13.3 to 35.7)	17	88.2	-	5.2	-
Diabetes¶								
yes	29	62.1	6.3 (-12.9 to 25.4)	5	100	9.1 (0.1 to ∞)§	∞	‖
no	172	55.8	-	77	90.9	-	6.1	
Smoking¶								
yes	114	60.5	8.8 (- 5.0 to 22.6)	31	93.6	3.6 (- 8.4 to 15.5)	9.4	1.8 (0.4- 9.0)
no	87	51.7	-	50	90.0	-	5.2	-
Beta-blocker use								
yes	123	57.7	0.6 (-13.1 to 14.3)	45	91.1	-	6.5	-
no	84	57.1	-	43	93.0	1.9 (- 9.3 to 13.2)	8.2	1.3 (0.3- 5.4)
Cholesterol (mmol/l)¶								
4.0- 6.0	52	51.9	-	33	87.9	-	4.3	-
6.1-12.0	150	61.3	9.4 (- 6.2 to 25.1)	48	93.8	5.9 (- 7.2 to 18.9)	9.9	2.3 (0.5- 9.9)
Expected load (Watt)¶								
70-149	94	50.0	-	45	93.8	3.8 (- 7.8 to 15.3)	8.1	1.3 (0.3- 5.3)
150-240	112	64.3	14.3 (0.8 to 27.7)†	43	90.0	-	6.4	-
SBP* baseline (mmHg)								
100-140	79	64.6	11.4 (-2.2 to 25.1)	52	96.2	10.1 (- 2.4 to 22.5)‡	17.0	4.5 (0.9-21.7)
141-240	128	53.1	-	36	86.1	-	3.8	-
Additional test variables								
Maximal load (Watt)								
45-134	162	62.4	22.4 (6.2 to 38.9)†	45	91.1	-	7.0	1.2 (0.3- 5.4)
135-280	45	40.0	-	43	93.0	1.9 (- 9.4 to 13.2)	5.7	-
Relative load (%)¶								
30- 90	154	66.2	33.5 (18.8 to 48.3)†	41	85.4	-	4.5	-
91-140	52	32.7	-	47	97.9	12.5 (0.9 to 24.1)‡	15.6	3.5 (0.4-28.1)
SBP* peak (mmHg)¶								
110-175	93	66.7	16.7 (3.1 to 30.2)†	42	92.9	2.0 (- 9.8 to 13.5)	9.4	1.7 (0.4- 7.3)
176-240	106	50.0	-	44	90.9	-	5.5	-
Disease specifications								
Number diseased vessels								
none				88	90.9			
one	71	39.4	-					
two	74	58.1	18.7 (2.7 to 34.7)†					
three	62	77.4	38.0 (22.6 to 53.4)†					

* Sens, sensitivity; RD, rate difference; Spec, specificity; CI, confidence interval; LR, likelihood ratio; LRR, likelihood ratio ratio; SBP, systolic blood pressure; - , reference category

† Determinant of sensitivity

‡ Determinant of specificity

§ Exact 95% CI of the odds ratio (95% CI of the RD could not be assessed because of 0 observations in one cell)

‖ Methodology could not be applied to an infinite likelihood ratio

¶ few values were missing