Diagnostic Research: improvements in design and analysis

Corné Biesheuvel

Diagnostic Research: improvements in design and analysis
Utrecht, Universiteit Utrecht, Faculteit Geneeskunde
Thesis, with a summary in Dutch
Proefschrift, met een samenvatting in het Nederlands

ISBN	90-393-2706-8
Author	C.J. Biesheuvel
Cover design & lay-out	J.C. Los
Print	Febodruk BV, Enschede

Diagnostic Research: improvements in design and analysis

Diagnostisch onderzoek: verbeteringen in studie opzet en analyse
(met een samenvatting in het Nederlands)

Proefschrift ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, Prof. Dr. W.H. Gispen ingevolge het besluit van het College voor Promoties in het openbaar te verdedigen op woensdag 27 april 2005 des middags te 14:30 uur

door
Cornelis Jan Biesheuvel
geboren op 3 april 1976 te Brakel

Promotores:

Prof. Dr. D.E. Grobbee
Julius Center for Health Sciences and Primary Care,
UMC Utrecht, The Netherlands

Prof. Dr. K.G.M. Moons
Julius Center for Health Sciences and Primary Care,
UMC Utrecht, The Netherlands

The research described in this thesis was funded by the Netherlands Organisation for Health Research and Development (ZonMw, # 904-66-112).

Financial support for the publication of this thesis by the Julius Center for Health Sciences and Primary Care and the Netherlands Organisation for Health Research and Development is gratefully acknowledged.

Additional support was received from Roche Diagnostics Nederland BV.

Manuscripts based on the studies presented in this thesis

Chapter 2
Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clin Chem 2004;50:473-476.

Chapter 3
Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomisation in diagnostic research. Submitted.

Chapter 4
Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Reappraisal of the nested case-control design in diagnostic research: updating the STARD guideline. Submitted.

Chapter 5
Biesheuvel CJ, Koomen I, Vergouwe Y, van Furth AM, Oostenbrink R, Moll HA, Grobbee DE, Moons KG. Validating and updating a prediction rule for neurological sequelae after childhood bacterial meningitis. Submitted.

Chapter 6
Biesheuvel CJ, Siccama I, Grobbee DE, Moons KG. Genetic programming or multivariable logistic regression in diagnostic research: a clinical example. J Clin Epidemiol 2004;57:551-660

Chapter 7
Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KG. Revisiting polytomous regression for diagnostic studies. Submitted.

Contents

chapter

1

Introduction

Diagnostic practice

To set a diagnosis is the cornerstone for medical care as it indicates treatment and provides an estimate of the patient's prognosis. A diagnostic test commonly has no direct therapeutic effects and does not directly influence patient outcome; setting a diagnosis is rather a vehicle to guide therapies. Once a diagnosis - or rather the probability of the most likely diagnosis - is established, and an assessment of the probable course of disease in the light of different treatment alternatives (including no treatment) has been made, a treatment decision has to be taken to eventually improve patient outcome.

In clinical practice, a diagnosis starts with a patient presenting with a particular set of symptoms or signs. The physician directly defines the possible diagnoses (i.e. differential diagnoses) and often implicitly determines the likelihood or probability of these diagnoses given the patient's symptoms or signs. The physician then determines the so-called target disease or working diagnosis to which the diagnostic work-up initially will be directed. This work-up commonly follows a phased approach, starting with patient history and physical examination. Subsequent steps may include additional tests such as laboratory tests, imaging, electrophysiology, biopsy, and angiography[1-4]. As long as uncertainty about the final diagnosis remains, further diagnostic tests are applied until a treatment decision can be made with sufficient confidence.

Hence, to set a diagnosis is a consecutive process of implicitly estimating the probability of disease presence or absence. Hardly any diagnosis is set by one test. Each test result, including the answer on a simple question like age or gender, is interpreted in view of other test results. Therefore, making a diagnosis is a multivariable concern. As different tests provide to varying extents the same information and each test may be more or less burdening to the patient, time consuming and costly, the true clinical relevance of a test is determined by its added or independent contribution to the probability estimation[2;5-11].

Diagnostic research

With diagnostic research we refer to scientific studies that aim to quantify whether and to what extent a (new) test additionally contributes to the estimation of presence or absence of a particular disease. The usual motive for diagnostic research is to improve the accuracy or to increase the efficiency (i.e. to decrease the patient burden and costs) of the current diagnostic work-up. To do so, diagnostic research should reflect the probabilistic character and multivariable work-up of diagnostic practice in design and analysis.

Diagnostic research, like prognostic research, is typically prediction research aiming to estimate absolute, rather than relative, disease probabilities. Prognostic research aims to predict the probability of future occurrence of a particular outcome in patients with a particular disease, whereas diagnostic research aims to predict or estimate the presence (or absence) of a particular disease in patients suspected of having that disease. Many diagnostic studies have resulted in so-called multivariable diagnostic prediction or decision rules. Such rules combine multiple test results and can be used in practice to estimate the probability of having a certain target disease for an individual patient. Well known examples are the Ottawa ankle rule to diagnose ankle fracture[12] and the Wells rule to diagnose deep venous thrombosis[13]. Given the commonly applied dichotomization in the diagnostic outcome (presence or absence of the target disease), such rules are usually developed with dichotomous logistic regression analysis[14-19]. Consequently, in such analysis the

1_ Feinstein AR.
Clinical Epidemiology: the architecture of clinical research.
Philadelphia: WB Saunders Company, 1985.

2_ Moons KG, Grobbee DE.
Diagnostic studies as multivariable, prediction research.
J Epidemiol Community Health 2002;56:337-8.

3_ Moons KG, Biesheuvel CJ, Grobbee DE.
Test research versus diagnostic research.
Clin Chem 2004;50:473-6.

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

5_ Begg CB.
Methodologic standards for diagnostic test assessment studies [editorial].
J Gen Intern Med 1988;3:518-20.

6_ Colditz GA.
Improving standards of medical and public health research.
J Epidemiol Community Health 2002;56:333-4.

7_ Feinstein AR.
Misguided efforts and future challenges for research on 'diagnostic tests'.
J Epidemiol Community Health 2002;56:330-2.

8_ Jaeschke R, Guyatt GH, Sackett DL.
Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients.
JAMA 1994;271:703-7.

9_ Knottnerus JA.
Challenges in dia-prognostic research.
J Epidemiol Community Health 2002;56:340-1.

alternative diseases of the differential diagnoses are included in the outcome category 'target disease absent'. However, it would be clinically more appealing if one could directly estimate the probability of presence of each of the diseases in the differential diagnoses for individual patients presenting with a particular symptom or sign. This requires diagnostic research analysing the polytomous diagnostic outcome, i.e. the differential diagnoses, rather than the dichotomous outcome.

Various reviews have demonstrated that the majority of published diagnostic accuracy studies still have methodological flaws in design and analysis or provide results with limited practical applicability[20-24]. This has been attributed to the absence of proper principles and methods for diagnostic research, in contrast to the strict guidelines that exist for therapeutic and etiologic studies. Apparently, there is still a gap between diagnostic research and diagnostic practice. This gap has been the motivation for the studies described in this thesis. This thesis describes various methods to improve the design and analysis of diagnostic research.

Outline of this thesis

This thesis consists of two parts. The first part addresses alternative methods for the design of diagnostic studies. Chapter 2 discusses why diagnostic studies follow a multivariable approach to assess the added value of a diagnostic test rather than a single test approach. The majority of published diagnostic studies are still single diagnostic test evaluations. The multivariable and probabilistic character of the diagnostic work-up is not reflected in the objective, design, analysis and presentation of the study. In chapter 3, we discuss the reasons why a randomised study design in diagnostic research is often not necessary to quantify whether a new test may (eventually) lead to improved patient outcome. Randomisation is used to properly study the preventive (etiologic) effect of a particular determinant, commonly a therapeutic intervention, on patient outcome. The use of a randomisation in diagnostic research changes an essential characteristic of diagnostic research; it turns prediction research into intervention or etiological research. In chapter 4, we elaborate on the use of the nested case-control design in diagnostic accuracy research. The case-control design has widely been disapproved for diagnostic research as it often yields biased results for the accuracy of the test(s) under study. In contrast, nested case-control studies include the cost effectiveness of case-control studies, but are not subject to the typical forms of bias that may occur in conventional case-control studies.

The second part of this thesis describes analytical methods for diagnostic research. Prediction rules tend to perform better on patients on which the rule has been developed than on new patients. In chapter 5, we validated a prediction rule for neurological sequelae after childhood bacterial meningitis. The rule was developed on a small sample of patients, and validated on a large sample of other children with bacterial meningitis, selected from almost all hospitals in The Netherlands. We tested the generalisability of the prediction rule and updated the rule after combining the derivation and validation sets. In chapter 6, we present genetic programming as an alternative for conventional dichotomous logistic regression analysis to develop diagnostic prediction rules. We compared the calibration and discrimination of a diagnostic rule derived by genetic programming to a rule derived by multivariable logistic regression analysis, using a data set comprising

10_ Sox H.
Probability theory in the use of diagnostic tests. An introduction to critical study of the literature.
Ann Intern Med 1986;104: 60-6.

11_ van der Schouw YT, Verbeek AL, Ruijs JH.
Guidelines for the assessment of new diagnostic tests.
Invest Radiol 1995;30:334-40.

12_ Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR.
A study to develop clinical decision rules for the use of radiography in acute ankle injuries.
Ann Emerg Med 1992;21: 384-90.

13_ Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C et al.
A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process.
J Intern Med 1998;243:15-23.

14_ Harrell 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.

15_ Knottnerus JA.
Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables.
Med Decis Making 1992;12: 93-108.

16_ Laupacis A, Sekar N, Stiell IG.
Clinical prediction rules. A review and suggested modifications of methodological standards.
JAMA 1997;277:488-94.

patients suspected of pulmonary embolism. In chapter 7, we examine and discuss the value of polytomous logistic regression using empirical data from a study on diagnosis of residual retroperitoneal mass histology in patients with nonseminomatous testicular germ cell tumour. We illustrate that a prediction rule derived by polytomous logistic regression may facilitate simultaneous prediction of the probabilities of presence of each of the differential diagnoses. Hence, this method may serve practice better than analysing a dichotomous outcome (target disease present, yes versus no) with conventional dichotomous logistic regression.

This thesis ends with concluding remarks on our findings. In addition, we provide suggestions for future research.

17_ Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.

18_ Knottnerus JA. Prediction rules: statistical reproducibility and clinical similarity. Med Decis Making 1992;12: 286-7.

19_ Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313:793-9.

20_ Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Meulen van der JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

21_ Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85-91.

22_ Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.

23_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

24_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7-18.

chapter

2

Test research versus diagnostic research

Introduction

The diagnostic work-up starts with a patient presenting with symptoms or signs suggestive of a particular disease. The work-up is commonly a consecutive process starting with medical history and physical examination and simple tests followed by more burdensome and costly diagnostic procedures. Generally, after each test all available results are converted (often implicitly) to a probability of disease, which in turn directs decisions for additional testing or initiation of appropriate treatment. Setting a diagnosis is a multitest or multivariable process of estimating and updating the diagnostic probability of disease presence given combinations of test results. Each test may be more or less burdensome to the patient, time-consuming, and costly. Different tests often provide to various degrees the same information because they are all associated with the same underlying disorder. Relevant for physicians is to know which tests are redundant and which have true, independent predictive value for the presence or absence of the target disease. Accordingly, studies of diagnostic accuracy should demonstrate which (subsequent) test results truly increase or decrease the probability of disease presence as estimated from the previous results, and to what extent.

Various reviews have demonstrated that the majority of published studies of diagnostic accuracy still have methodological flaws in design or analysis or provide results with limited practical applicability[1-3]. This has been attributed to the absence of a proper principles and methods for diagnostic test evaluations as, for example, exists for studies of therapies and etiologic factors and has motivated various researchers to establish guidelines for studies of diagnostic accuracy, such as the recent STARD initiative[4-12]. In our view, an issue that has received too little attention in most of these methodological essays is the difference between test research and diagnostic research.

With 'test research' we refer to studies that follow a single-test or univariable approach, i.e., studies focusing on a particular test to quantify its sensitivity, specificity, likelihood ratio (LR), or area under the ROC curve (ROC area). We call this test research because it merely quantifies the characteristics of the test rather than the test's contribution to estimate the diagnostic probability of disease presence or absence. By 'diagnostic research' we refer to studies that aim to quantify a test's added contribution beyond test results readily available to the physician in determining the presence or absence of a particular disease. Although the multivariable and probabilistic character of medical diagnosis is slowly gaining appreciation in medical research, the majority of studies on diagnostic accuracy may still be regarded as test research[1;2;10].

We believe that test research has limited applicability to clinical practice. Below we describe why we believe this is the case, provide a brief description of a better approach, and give two clinical examples illustrating the hazards of test research. Finally, we describe the few instances in which test research may be worthwhile.

The first reason that test research has limited relevance to practice is the nature of the questions that are usually addressed. The practical utility of the estimation of sensitivity, specificity, and LR for a particular test in the diagnosis of a particular disease is not always obvious[11;13]. Consider, for example, the diagnostic workup for patients suspected of deep vein thrombosis (DVT). The relevant research question for patients suspected of DVT would be: "Given patient history and physical examination, which subsequent tests (e.g., D-dimer measurement) truly provide added information to predict the presence or absence of DVT?" The probability of disease presence and quantifying which tests independently contribute to the

1_ Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Meulen van der JH et al.
Empirical evidence of design-related bias in studies of diagnostic tests.
JAMA 1999;282:1061-6.

2_ Mower WR.
Evaluating bias and variability in diagnostic test reports.
Ann Emerg Med 1999;33:85-91.

3_ Reid MC, Lachs MS, Feinstein AR.
Use of methodological standards in diagnostic test research. Getting better but still not good.
JAMA 1995;274:645-51.

4_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al.
Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative.
BMJ 2003;326:41-4.

5_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al.
The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration.
Clin Chem 2003;49:7-18.

6_ Fryback D, Thornbury J.
The efficacy of diagnostic imaging.
Med Decis Making 1991;11:88-94.

7_ Hunink MG, Krestin GP.
Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology.
Radiology 2002;222:604-14.

8_ Jaeschke R, Guyatt GH, Sackett DL.
Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid?
JAMA 1994;271:389-91.

estimation of this probability should be the objects of study. However, in this respect many studies have aimed only to estimate the sensitivity and specificity of the D-dimer assay. When this is the object of a study, it is only the probability of obtaining a positive or negative test result that is addressed, rather than the probability of disease presence. Moreover, the focus is on the value of a single test rather than on the value of that test in combination with other, previous tests, including patient history and physical examination. We may say that the object of research is the test rather than the (probability of) disease. Hence the term test research.

Test characteristics are not fixed

The second reason that results from test research have limited relevance is that a test's sensitivity, specificity, LR, and ROC area tend to be taken as properties or characteristics of a test. This, however, is a misconception, as we discussed recently[13]. It is widely accepted that the predictive values of a test vary across patient populations. However, several studies have empirically shown that the sensitivity, specificity, and LR of a test may vary markedly, not only across patient populations[14] but also within a particular study population[13;15-17]. Within different patient subgroups, defined by patient characteristics or other test results, a particular test may have different sensitivities and specificities. This is because all diagnostic results obtained from patient history, physical examination, and additional tests are to some extent related to the same underlying disorder. For example, immobility, gender, and use of oral contraceptives are associated with the development of DVT. In turn, the presence of DVT determines the presence of symptoms and signs and also (the probability of finding) a positive D-dimer assay result. Accordingly, via the underlying disorder, all diagnostic results are somehow correlated and thus mutually determine each other's sensitivity, specificity, and LR to various extents[13;15-17]. A single value of a test's sensitivity, specificity, LR, ROC area, or predictive value that applies to all patients of a study sample does not exist. Hence, there are no fixed test characteristics.

Selection bias

The most widely acknowledged limitation of test research is that studies often apply an improper patient recruitment and study design[1-3;11]. Investigators often select study participants among those who underwent the reference test in routine practice, i.e., selection based on a 'true' presence or absence of the disease. The results of the test(s) under study are retrieved from the medical records and then compared across those with and without the disease. Such a case-control design commonly leads to selection bias, known as verification, work-up, or referral bias[9;18;19]. Although such patient recruitment methods and study designs have decreased in the past decade, test research is still frequently based on individuals selected based on their final diagnosis[1-3]. The need for proper patient recruitment is extensively addressed in the STARD checklist[4;5]. Study participants should be selected in agreement with the indication for diagnostic testing in practice, i.e., on their suspicion of having a particular disease, rather than on the presence or absence of that disease. Such unbiased selection of study participants may indeed be problematic for diagnostic laboratories or imaging centers that do not have access to consecutive series of patients suspected of having the disease. Moreover, most hospital databases code patients according to their final diagnosis rather than by their presenting symptoms or signs. The use of a system to register patients not only on their final diagnosis but also on their clinical presentation

9_ Knottnerus JA.
The evidence base of clinical diagnosis.
London: BMJ Publishing group, 2002.

10_ Moons KG, Grobbee DE.
Diagnostic studies as multivariable, prediction research.
J Epidemiol Community Health 2002;56:337-8.

11_ Sackett DL, Haynes RB.
The architecture of diagnostic research.
BMJ 2002;324:539-41.

12_ van der Schouw YT, Verbeek AL, Ruijs JH.
Guidelines for the assessment of new diagnostic tests.
Invest Radiol 1995;30:334-40.

13_ Moons KG, Harrell FE.
Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies.
Acad Radiol 2003;10:670-2.

14_ Fletcher RH.
Carcinoembryonic antigen.
Ann Intern Med 986;104:66-73.

15_ Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA.
Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis.
Am J Med 1984;77:64-71.

16_ Levy D, Labib SB, Anderson KM, Christiansen JC, Kanell WB, Castelli WP.
Determinants of sensitivity and specificity of electrocardiagraphic criteria for left ventricular hypertrophy.
Circulation 1990;81:815-20.

17_ Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE.
Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example.
Epidemiology 1997;8:12-7.

would enhance the validity and clinical relevance of diagnostic accuracy research[20].

We believe that to serve practice, the point of departure and the multivariable and probabilistic character of the diagnostic workup should be reflected in the objective, design, analysis, and presentation of studies of diagnostic accuracy. The aim is to relate the probability of disease presence to combinations of test results, following their typical chronology in practice. The predictive accuracy of the initial tests (including patient history and physical examination) should be estimated first, and the added value of more burdening and costly tests should be estimated subsequently. Hence, all tests typically applied in the workup need to be documented in each patient, even if a study focuses on a particular test. Consider again the question whether the D-dimer assay is relevant to the diagnosis of DVT. A consecutive series of patients suspected of DVT should be selected. The history, physical examination, and D-dimer result should be obtained from each patient. Subsequently, each patient "undergoes" the best reference test currently available; in this example, it would be repeated leg ultrasound. What to do in the absence of a single reference test or when it is unethical to perform the reference test in each patient has been described elsewhere[7;10;21;22].

Because the D-dimer assay will always be applied after history taking and physical examination, the statistical analysis requires a comparison of the (average) probability of disease presence without and with the D-dimer assay, overall or in subgroups. Such sequential modelling of the diagnostic probability as a function of different combinations of test results can be done using, e.g., multivariable logistic regression. Such multivariable analyses account for the mutual dependencies between different test results and thus indicate which tests truly do and which do not independently contribute to the estimation of the probability of disease presence. In addition, various orders of diagnostic testing can be analysed. The result of such analysis is the definition of one or more diagnostic prediction models including only the relevant tests. If needed, such prediction models can be simplified to obtain readily applicable diagnostic decision rules for use in practice. Various authors have applied or described the details of such an analytical approach[20;23-27].

Multivariable diagnostic prediction models or rules are not the solution to everything. They may have several drawbacks, such as overoptimism, although methods have been described to overcome some of these drawbacks[23]. The need for multivariable modelling in diagnostic research, however, is not different from other types of medical research, such as etiologic, prognostic, and therapeutic research. It is not the singular association between a particular exposure or predictor and the outcome that is informative, but their association independent of other factors. For example, in etiologic research, investigators never publish the crude estimate between exposure and outcome only, but always the association in view of other risk factors (confounders), using a multivariable analysis as well[13]. Similarly, in diagnostic accuracy research, multivariable modelling is necessary to estimate the value of a particular test in view of other test results. As in other types of research, such knowledge cannot be inferred from singular, univariable test parameters[10;11;13].

Fortunately, a multivariable approach in design and analysis aiming to quantify the independent value of diagnostic tests has gained approval[20;23-27]. In addition, the above study question on the added value of the D-dimer assay in diagnosing DVT has been evaluated in such a way. The D-dimer assay appeared to have an added predictive value to patient history and physical examination, particularly in patients who have a low clinical probability of DVT[27].

18_ Begg CB, Greenes RA.
Assessment of diagnostic tests when disease verification is subject to selection bias.
Biometrics 1983;39:297-15.

19_ Ransohoff DF, Feinstein AR.
Problems of spectrum and bias in evaluating the efficacy of diagnostic tests.
N Engl J Med 1978;299:926-30.

20_ Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE.
Diagnostic research on routine care data: prospects and problems.
J Clin Epidemiol 2003;56:501-6.

21_ Moons KG, Grobbee DE.
When should we remain blind and when should our eyes remain open in diagnostic research.
J Clin Epidemiol 2002;55:633-6.

22_ Bossuyt PM, Lijmer JG, Mol BW.
Randomised comparisons of medical tests: sometimes invalid, not always efficient.
Lancet 2000;356:1844-7.

23_ Harrell FE.
Regression modeling strategies.
New York: Springer-Verlag, 2001.

24_ Laupacis A, Sekar N, Stiell IG.
Clinical prediction rules. A review and suggested modifications of methodological standards.
JAMA 1997;277:488-94.

25_ Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE.
Redundancy of single diagnostic test evaluation.
Epidemiology 1999;10:276-81.

We now present two clinical examples illustrating how results from a single or univariable test approach can mislead.

In an Australian study, 399 consecutive dyspeptic patients referred for endoscopy underwent two tests, the rapid urease test and the 13C breath test, for *Helicobacter pylori* (HP) with endoscopy as the reference test[28]. The investigators found large differences in the test results between patients with a normal and abnormal endoscopy. The sensitivity and specificity were 96% and 67% for the rapid urease test and 91% and 82% for the 13C breath test. The authors concluded that the HP tests might have potential for the initial evaluation of dyspepsia and needed further evaluation in general practice. A second study was done by Weijnen et al.[26] Using a sequential multivariable approach, they found in a consecutive series of 565 dyspeptic patients referred for endoscopy that the HP test did not add diagnostic information to the predictors from history (i.e., history of ulcer, pain on empty stomach, and smoking). The ROC area of the model with only predictors from patient history was 0.71, which was increased to only 0.75 ($p = 0.46$) after addition of the HP test result. They concluded that HP testing in all dyspeptic patients has no value in addition to history taking.

Cowie et al.[29] studied a consecutive series of 122 patients suspected of heart failure. They measured in each patient the plasma concentrations of three natriuretic peptides, A-type natriuretic peptide (ANP), N-terminal ANP, and B-type natriuretic peptide (BNP), as well as the presence or absence of heart failure, using consensus diagnosis based on chest radiography and echocardiography as the reference test. They found that the mean concentration of each natriuretic peptide separately (single test approach) was significantly greater in the patients with heart failure (all $P < 0.001$). They also evaluated all three together in a multivariable logistic prediction model. Only the BNP measurement remained significantly associated with heart failure presence, whereas the other two did not add any predictive information.

Both examples show that one may qualify a test differently (commonly more promisingly) when only the results of a univariable or single test approach are considered. Evaluating a particular test in view of other test results and accounting for mutual dependencies may decrease or even diminish its diagnostic contribution, simply because the information provided by that test is already provided by the other tests. Because in real life any test result is always considered in view of other patient characteristics and test results, diagnostic accuracy studies that address only a particular test and its characteristics have, in our view, limited relevance to practice. Indeed, as shown by Reid et al.[30], test characteristics are hardly ever actually used by practitioners.

There are two situations in which pure test research, i.e., studies aiming to estimate the diagnostic accuracy indices of a single test, is indicated. The first situation is when a diagnosis is indeed set by only one test and other test results are not considered. This is, in our view, reserved to the context of screening for preclinical stages of a particular disease: e.g., screening for breast cancer, prostate cancer, or cervical cancer. Such screening may be considered as a specific case of diagnosis, concerned with the early detection of a disease in a particular age and sex group. Here, only the screening test is considered in the diagnostic process; other patient characteristics or test results are commonly not available and therefore cannot modify the sensitivity, specificity, LR, and predictive values of the screening test. Accordingly, these indices, as estimated from a particular study sample, may be

26_ Weijnen CF, Numans ME, de Wit NJ, Smout AJ, Moons KG, Verheij TJ et al.
Testing for Helicobacter pylori in dyspeptic patients suspected of peptic ulcer disease in primary care: cross sectional study.
BMJ 2001;323:71-5.

27_ Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L et al.
Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis.
Thromb Haemost 1999;81:493-7.

28_ Fraser AG, Ali MR, McCullough S, Yeates NJ, Haystead A.
Diagnostic tests for Helicobacter pylori--can they help select patients for endoscopy?
N Z Med J 1996;109:95-8.

29_ Cowie MR, Struthers AD, Wood DA, Coats AJ, Thompson SG, Poole-Wilson PA et al.
Value of natriuretic peptides in assessment of patients with possible new heart failure in primary care.
Lancet 1997;350:1349-53.

30_ Reid MC, Lane DA, Feinstein AR.
Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy.
Am J Med 1998;104:374-80.

considered characteristics or constants for the corresponding source population. In the presence of a positive screening result, patients are commonly referred for further diagnostic workup. Other test results then become involved, and mutual dependencies between the screening test and these other tests start to play a role, demanding a multivariable approach in design and analysis.

The second situation, as suggested previously, is in the initial phase of developing a new test or evaluating an existing test in a new context; single test evaluations in these circumstances may be useful for efficiency reasons[6;11;12;25]. Such initial test research should apply a case–control approach, preferably starting with a sample of patients with the disease (cases) and a sample of healthy controls. If the test cannot differentiate between these two extreme or heterogeneous outcome categories, the test development process would likely be terminated. In such instances, it will be unlikely that the test does show discriminative value in patients suspected of having the disease, i.e., the population for which the test is intended, because these patients present with similar disease profiles, leading to an even more homogeneous case mixture. However, once the test does yield 'satisfactory' diagnostic indices in such an initial test research study, we believe that its independent predictive contribution to existing diagnostic information in a clinical context can and must still be quantified by the above proposed approach.

chapter

3

Distraction from randomisation in diagnostic research

Introduction

In almost every system to grade epidemiological studies according to their level of evidence, randomised studies or meta-analyses of randomised studies receive the highest classification[1-4]. Although the use of such hierarchies may help to separate the wheat from the chaff, it has also led to misconception and abuse [5-8]. The paradigm of a randomised study has also been applied to diagnostic research questions[9-13].

The ultimate goal of diagnostic testing is, like all medical care, to improve patient outcome. Hence, it has widely been advocated that when establishing a test's diagnostic accuracy, the impact of the test on patient outcome must also be quantified[9;10;12;14;15]. However, to demonstrate the beneficial effect of a diagnostic procedure or strategy on patient outcome, we believe that randomisation is by no means a prerequisite. The use of randomisation will transform a diagnostic study or test evaluation into an etiologic or intervention study which may not be necessary in many instances. The nature of the diagnostic question and the best way to find empirical evidence to answer this question determines the appropriate study design[5;6].

Diagnostic practice without randomisation

Setting a diagnosis in a patient suspected of a particular disease is to estimate the probability of the presence or absence of this disease, based on the diagnostic information obtained from patient history, physical examination plus additional testing[16-20]. Setting a diagnosis in itself is not a therapeutic process, but rather a vehicle to guide therapies. Moreover, a diagnostic test commonly has no direct therapeutic effects and therefore does not directly influence patient outcome. Once a diagnosis or rather the probability of the most likely diagnosis is established and an assessment of the probable course of disease in the light of different treatment alternatives including no treatment has been made, a treatment decision has to be taken to eventually improve patient outcome.

Diagnostic research

Most diagnostic test evaluations include diagnostic accuracy research. Diagnostic accuracy research typically involves cross-sectional studies quantifying the accuracy of a new or existing diagnostic test as compared to a reference test or method which is the best method available at the start of the study[17;21;22]. There is no patient follow-up and patient outcome is not considered. The test under study is evaluated on its predictive accuracy, i.e. its ability to discriminate between the 'true' presence and absence of the disease or, otherwise, to properly estimate the probability of presence of that disease. The aim of accuracy studies is to investigate whether the usually more burdening, time consuming or costly reference test can be replaced by the test under study which is commonly less invasive, time consuming or costly. Such replacement is indicated if the test under study is as accurate as the reference method, i.e. produces similar diagnostic classifications or at least at acceptable percentages of incorrect classifications.

If a cross-sectional diagnostic accuracy study has indicated that the test(s) under study indeed may appropriately replace the existing reference and thus similarly classify the presence or absence of the disease under study, the effect of that test(s) on patient outcome can be validly established without the need of a randomised follow-up study. When other properly executed therapeutic studies

1_ Trout KS.
How to read clinical journals: IV. To determine etiology or causation.
Can Med Assoc J 1981;124: 985-90.

2_ Sacks H, Chalmers TC, Smith H, Jr.
Randomized versus historical controls for clinical trials.
Am J Med 1982;72:233-40.

3_ Horwitz RI, Feinstein AR.
Methodologic standards and contradictory results in case-control research.
Am J Med 1979;66:556-64.

4_ Pocock SJ, Elbourne DR.
Randomized trials or observational tribulations?
N.Engl.J.Med. 2000;342: 1907-9.

5_ Glasziou P P.
Vandenbroucke J, Chalmers I.
Assessing the quality of research.
BMJ 2004;328:39-41.

6_ Sackett DL, Wennberg JE.
Choosing the best research design for each question.
BMJ 1997;315:1636.

7_ Black N.
Why we need observational studies to evaluate the effectiveness of health care.
BMJ 1996;312:1215-8.

8_ McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C.
Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies.
BMJ 1999;319:312-5.

9_ Dixon AK.
Evidence-based diagnostic radiology.
Lancet 1997;350:509-12.

10_ Fryback D, Thornbury J.
The efficacy of diagnostic imaging.
Med Decis Making 1991;11: 88-94.

have quantified the occurrence of patient outcome given the different treatment possibilities of that disease, one could quantify the value of the test to improve patient outcome by combining the results of the cross-sectional diagnostic accuracy studies and of the longitudinal randomised therapeutic studies using simple statistical or decision modelling techniques[23;24].

Hence, in our view, a test's effect on patient outcome can be inferred and indeed considered as quantified 1) if the test is meant to include or exclude a disease for which an established reference is available, 2) if a cross-sectional accuracy study has shown the test's ability to adequately detect the presence or absence of that disease based on the reference, and finally 3) if proper, randomised therapeutic studies have provided evidence on efficacy of the optimal management of this disease[9;10;22;25]. In such instances diagnostic research does not require an additional randomised comparison between two (or more) test-treatment strategies, with and without the test under study, to establish the test's effect on patient outcome.

An example of diagnostic accuracy research in which a (randomised) therapeutic study was not necessary to quantify the effect of a new test on patient outcome is given by a study in which a new immunoassay for the detection of *Helicobacter pylori* infection was compared with the established reference (a combination of rapid urease test, urea breath test and histology) in a primary care setting[26]. This study demonstrated that the new diagnostic test could substitute the more costly and invasive reference test (histology which requires endoscopy). As therapeutic management, established by randomised clinical trials[27], of patients infected with *H. pylori* remained unaltered, another randomised study to quantify the effect of the use of the new immunoassay test on patient outcome was not needed.

Randomisation in diagnostic research

Randomisation is particularly common in causal or etiologic studies aiming to quantify the intended effect of determinants (notably preventive or therapeutic strategies) on patient outcome[28]. In this, randomisation ensures - provided large enough number of subjects and properly executed - similar distribution of the other causal factors or confounders across the categories of the interventions under study. In this case, any observed difference in patient outcome between both study groups can be attributed to the intervention studied. This intervention can be a single intervention, a combination of interventions or a test-treatment intervention. As explained earlier, in diagnostic research randomised studies are commonly no prerequisite to validly quantify the value of a diagnostic test on patient outcome. We believe that (follow-up) studies using a randomised design to quantify the value of a diagnostic test on patient outcome are only indicated if:

1. The disease at issue has an imperfect reference test, such as depression, irritable bowel syndrome, and congestive heart failure;
2. The diagnostic technology under study might be better to the extent that it provides new information, potentially leading to other treatment choices than the existing reference. For example in functional imaging with positron emission tomography (PET) in diagnosing pancreatic cancer, for which computed tomography (CT) is the current reference;
3. There is no direct link between the result of the new diagnostic test under study and an established treatment indication, such as the finding of non-calcified small nodules (less than 5.0 mm) when screening for lung cancer with low-dose spiral CT scanning[29];

11_ Hunink MG, Krestin GP.
Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology.
Radiology 2002;222:604-14.

12_ Mackenzie R, Dixon AK.
Measuring the effects of imaging: an evaluative framework.
Clin Radiol 1995;50:513-18.

13_ Bossuyt PM, Lijmer JG, Mol BW.
Randomised comparisons of medical tests: sometimes invalid, not always efficient.
Lancet 2000;356:1844-7.

14_ Freedman LS.
Evaluating and comparing imaging techniques: a review and classification of study designs.
Br J Radiol 1987;60:1071-81.

15_ Hunink MG.
Outcomes research and cost-effectiveness analysis in radiology.
Eur Radiol 1996;6:615-20.

16_ Moons KG, Harrell FE.
Sensitivity and specificity should be deemphasized in diagnostic accuracy studies.
Acad Radiol 2003;10:670-2.

17_ Moons KG, Biesheuvel CJ, Grobbee DE.
Test research versus diagnostic research.
Clin Chem 2004;50:473-6.

18_ Feinstein AR.
Clinical Epidemiology: the architecture of clinical research.
Philadelphia: WB Saunders Company, 1985.

19_ Sackett DL, Haynes RB, Tugwell P.
Clinical epidemiology; a basic science for clinical medicine.
Boston: Little, Brown & Co, 1985.

4. The diagnostic technology under study in itself may directly have therapeutic properties such as salpingography to determine patency of the uteral tubes[30].

Ad 1. Ideally, in diagnostic accuracy studies the final diagnosis is made by the reference method without knowledge of the results of the test(s) under study. Diagnostic accuracy studies of diseases with an imperfect reference test commonly use a consensus diagnosis as reference method[19;21;31;32]. In this, an independent expert panel assigns a final diagnosis to each patient often using all available patient information, including the diagnostic information of the test under study. The use of such reference, however, inherently carries the danger of so-called incorporation bias, in which the results of the test under study are incorporated in the assessment of the final diagnosis[31;33;34]. This commonly leads to overestimation of the test under study. However, withholding the results of the (new) test(s) under study may lead to misclassification of the final diagnosis with varying consequences (over- or underestimation of the test's accuracy) that can hardly be judged afterwards. There are no general solutions to this dilemma, which is inherent to studying diseases that lack a proper reference method to determine their presence or absence[31;33;34].

If the disease lacks an established reference method, the only way to fully prevent this bias and validly quantify a test's diagnostic value to improve patient outcome, is to directly study the test's contribution on patient outcome. For example, one could perform a randomised follow-up study, comparing the test-treatment strategy with the test under study to the test-treatment-strategy without the test under study[11;13]. Only such randomised study design can provide evidence for the (added) value of the test in the diagnosis and treatment of a disease that lacks an established reference.

Ad 2. An example of a new test that might provide better or other information, potentially leading to other treatment choices than the existing reference is functional imaging with PET in diagnosing pancreatic cancer[35]. Compared to CT, PET may especially be helpful in detecting smaller lesions and distant metastases[36;37]. Application of PET may thus lead to other diagnostic classifications and thus initiating other treatment choices potentially leading to different patient outcomes than the use of CT. A simple diagnostic accuracy study comparing PET with CT as reference is not sufficient to quantify afterwards the true effects on patient outcome when using PET in the diagnosis of pancreatic cancer, as the potential treatments might be different. Preferably, one should perform a randomised follow-up study on patient outcome, comparing both diagnostic tests with corresponding treatment choices. In this there are several possibilities to randomise the patients suspected of pancreatic cancer[11;13].

Ad 3. Presently, the finding of non-calcified nodules less than 5.0 mm on CT during screening for lung cancer, does not directly indicate a particular treatment[29]. A period of observation by CT is the initial mode of action in the presence of non-calcified small nodules[38]. A cross-sectional diagnostic accuracy study comparing CT with chest radiography as reference would not be sufficient to quantify the true effects of CT in screening for lung cancer on patient outcome, as application of CT may lead to other treatment choices potentially leading to different patient outcomes than the use of radiography. Several randomised and non-randomised follow-up studies of the effect of screening for lung cancer on patient outcome have been started, to study the effect of CT scanning on patient outcome[31;39-41].

20_ Grobbee DE, Miettinen OS. Clinical Epidemiology: introduction to the discipline. Neth J Med 1995;47:2-5.

21_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

22_ Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337-8.

23_ Hunink MG, Glasziou P. Decision making in health and medicine. 2001. Cambridge, United Kingdom, Cambridge University Press.

24_ Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: WB Saunders Company, 1980.

25_ Moons KG, Ackerstaff RGA, Moll FL, Spencer MP, Algra A. Association of intraoperative transcranial doppler monitoring variables with stroke from carotid endarterectomy (respons). Stroke 2001;32:813.

26_ Weijnen CF, Hendriks HA, Hoes AW, Verweij WM, Verheij TJ, de Wit NJ. New immunoassay for the detection of Helicobacter pylori infection compared with urease test, 13C breath test and histology: validation in the primary care setting. J Microbiol Methods 2001;46:235-40.

Ad 4. Although rare, a diagnostic test in itself may directly have therapeutic properties, as for example is the case for salpingography to determine patency of the uteral tubes in women suspected of uteral tube obstruction. In these instances, the test does not only serve as a diagnostic tool to guide therapeutic decisions that in turn may affect patient outcome, but as a therapeutic intervention as well. Properly quantifying such test's contribution on patient outcome obviously requires a randomised follow-up study, comparing patient outcome among patients diagnosed and treated with the use of the test compared to patients diagnosed and treated without the test under study.

In all above situations, a randomised study is the paradigm to allow for a valid estimate of the test's effect on patient outcome. Previous reports have already elaborated on the most efficient method of such randomised design as well as at which point in time patients should be randomised[11;13]. However, a limitation of a randomised approach to directly quantify the contribution of a diagnostic test on patient outcome is that it inherently addresses diagnosis and treatment as one combined strategy - a 'package deal'. Due to this, it is not possible to determine whether a positive effect on patient outcome can be attributed solely to the improved diagnosis or to the impact of other chosen treatment strategies[11;13;25].

27_ McColl KE, Murray LS, Gillen D, Walker A, Wirz A, Fletcher J et al.
Randomised trial of endoscopy with testing for Helicobacter pylori compared with non-invasive H pylori testing alone in the management of dyspepsia.
BMJ 2002;324:999-1002.

28_ Miettinen OS.
The need for randomization in the study of intended effects.
Stat Med 1983;2:267-71.

29_ Henschke CI, Yankelevitz DF, Naidich DP, McCauley DI, McGuinness G, Libby DM et al.
CT Screening for Lung Cancer: Suspiciousness of Nodules according to Size on Baseline Scans.
Radiology 2004.

30_ Papaioannou S, Afnan M, Girling AJ, Ola B, Olufowobi O, Coomarasamy A et al.
Diagnostic and therapeutic value of selective salpingography and tubal catheterization in an unselected infertile population.
Fertil Steril 2003;79:613-7.

31_ Moons KG, Grobbee DE.
When should we remain blind and when should our eyes remain open in diagnostic research.
J Clin Epidemiol 2002;55: 633-6.

32_ Knottnerus JA, Muris JW.
Assessment of the accuracy of diagnostic tests: the cross-sectional study.
J Clin Epidemiol 2003;56: 1118-28.

33_ Swets JA. Measuring
The accuracy of diagnostic systems.
Science 1988;240:1285-93.

34_ Weller SC, Mann PC.
Assessing rater performance without a "gold standard" using consensus theory.
Med Decis Making 1997;17:71-9.

Conclusions

Various research methods have their particular advantages and disadvantages, and the popular belief that only randomised studies produce results applicable to clinical practice with confidence and that observational studies may always be misleading does a disservice to patient care, clinical investigation and the education of health care professionals[5;6;42]. In many instances, randomised studies in diagnostic research are not necessary and cross-sectional accuracy studies are fully acceptable to validly estimate the value of the diagnostic test in improvements of patient care.

35_ Tamm EP, Silverman PM, Charnsangavej C, Evans DB.
Diagnosis, staging, and surveillance of pancreatic cancer.
Am J Roentgenol 2003;180: 1311-23.

36_ Hanbidge AE.
Cancer of the pancreas: the best image for early detection: CT, MRI, PET or US?
Can J Gastroenterol. 2002; 16:101-5.

37_ Saisho H, Yamaguchi T.
Diagnostic imaging for pancreatic cancer: computed tomography, magnetic resonance imaging, and positron emission tomography.
Pancreas 2004;28:273-8.

38_ Libby DM, Smith JP, Altorki NK, Pasmantier MW, Yankelevitz D, Henschke CI.
Managing the Small Pulmonary Nodule Discovered by CT.
Chest 2004;125:1522-9.

39_ Henschke CI, Naidich DP, Yankelevitz DF, McGuinness G, McCauley DI, Smith JP et al.
Early lung cancer action project: initial findings on repeat screenings.
Cancer 2001;92:153-9.

40_ Swensen SJ, Jett JR, Hartman TE, Midthun DE, Sloan JA, Sykes AM et al.
Lung cancer screening with CT: Mayo Clinic experience.
Radiology 2003;226:756-61.

41_ Patz EF, Swensen SJ, Herndon JE.
Estimate of lung cancer mortality from low-dose spiral computed tomography screening trials: implications for current mass screening recommendations.
J Clin Oncol 2004;22:2202-6.

42_ Concato J, Shah N, Horwitz RI.
Randomized, controlled trials, observational studies, and the hierarchy of research designs.
N Engl J Med 2000;342: 1887-92.

chapter
4

Reappraisal of the nested case-control design in diagnostic research: updating the STARD guideline

Introduction

To set a diagnosis in medical practice implies estimation of the absolute -rather than relative- probabilities of disease presence given the combination of test results, documented from the patient. Diagnostic studies commonly aim to quantify to what extent results of a particular (new) test predict the presence or absence of a particular disease in comparison to a reference method (diagnostic outcome) in patients suspected of that disease. This reference method can be a single test, a combination of tests or a consensus diagnosis. It is advocated that diagnostic studies apply a cross-sectional approach[1-7]. A cohort of patients with an indication for the diagnostic procedure at interest is selected, defined by the patients' suspicion of having the disease of interest. All patients will undergo the test(s) under study and subsequently the reference method. In the analysis, one can estimate measures of diagnostic accuracy such as sensitivity, specificity, likelihood ratios, diagnostic odds ratio, receiver operating characteristic (ROC) curve, and most importantly, the absolute probabilities of disease presence per (combination of) test result(s), i.e. the predictive values.

As an alternative to the cohort approach, a case-control approach may be used. In this design, patients are selected on the 'true' presence or absence of the disease under study, based on the reference test[3;8-10]. However, the use of a case-control design in diagnostic research has widely been disapproved[2-6;8-12]. An important disadvantage of a diagnostic case-control study is that this design may yield biased estimates of the diagnostic accuracy of the test(s) under study because of so-called verification (or work-up or referral) bias[1;3;8-10;12]. Physicians selectively refer patients for additional tests including the reference test based on previous test results whereas ideally all subjects suspected of the disease should undergo the reference method irrespective of a more or less prominent indication. A second disadvantage is that absolute probabilities of disease presence by test result (predictive values) cannot be obtained. This is because cases and controls are sampled from a source population, i.e. the patients suspected of the disease at interest, with unknown sample size. Therefore the sampling fraction of controls is unknown and a valid estimate of the predictive value cannot be calculated.

Recent meta-analyses on design-related bias in studies of diagnostic tests showed that the case-control design is still often used in diagnostic studies[9-11]. Accordingly, the recent Standards for Reporting of Diagnostic Accuracy (STARD) guideline emphasised to select study subjects on their suspicion of disease, i.e. not to use a case-control approach[5;6]. In spite of these well accepted limitations of the case-control design in diagnostic research, however, in our view an exception should be made for the nested case-control approach. In this paper we discuss why the nested case-control design can be a valid and efficient alternative for the traditional cohort approach. This is illustrated with empirical data of a cross-sectional cohort study on diagnosis of deep vein thrombosis (DVT).

1_ Begg CB, McNeill BJ.
Assessment of radiologic tests; control of bias and other design considerations.
Radiology 1988;167:565-9.

2_ Jaeschke R, Guyatt GH, Sackett DL.
Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid?
JAMA 1994;271:389-91.

3_ van der Schouw YT, van Dijk R, Verbeek AL.
Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests.
J Clin Epidemiol 1995;48: 417-22.

4_ Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE.
Diagnostic research on routine care data: prospects and problems.
J Clin Epidemiol 2003;56: 501-6.

5_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al.
Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative.
BMJ 2003;326:41-4.

6_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al.
The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration.
Clin Chem 2003;49:7-18.

7_ Moons KG, Biesheuvel CJ, Grobbee DE.
Test research versus diagnostic research.
Clin Chem 2004;50:473-6.

The nested case-control design in diagnostic research

The essential difference between a conventional case-control design and a nested case-control design is that in the former the source population is sampled but its size is not known, whereas a nested case-control study is 'nested' in an existing predefined cohort with known sample size[13;14]. In diagnostic research, a nested case-control study includes patients suspected of having the disease. The 'true' disease status is obtained for all cohort members using a reference method. Hence, there is no selection or work-up or verification bias. Typically, the results of the tests under study are retrieved or obtained for all subjects with the disease but only for a sample of the patients without the disease. All measures of diagnostic accuracy, including predictive values, can simply be obtained by weighing the controls by the case-control sample fraction, as explained in Figure 1.

To answer diagnostic questions, the nested case-control design is generally more efficient than a full cohort approach and, when conducted appropriately, equally valid. The efficiency of the approach may be particularly beneficial when the test under study is costly. This may for example apply to the measurement of genetic markers, tumour markers or neurohormones in blood samples. With a nested case-control approach only the blood of a fraction of the initial cohort needs to be analysed. Likewise, when the test under study includes the results of ECG analysis or imaging tests, the ECGs and images are obtained in all cohort members. But they need to be interpreted by a cardiologist or radiologist in only a fraction of the cohort. Finally, the nested case-control design may be particularly attractive in diagnostic research that requires the re-analysis of data from an existing cohort in which biological material (e.g. blood samples) has been stored for all patients. For example, when a new biological marker for a particular disease is discovered, one could retrospectively analyse the biological material of only a fraction of the total cohort, and compare the test results to the true presence or absence of the disease at interest without having to perform a new cohort study from the start.

In our view, results from diagnostic nested case-control studies -after weighing the controls by the sample fraction- should be virtually identical to results based on a full cohort analysis. To empirically document the validity of the nested case-control design in diagnostic research, we use data of a diagnostic cohort study, comprising 1295 patients who were suspected of DVT and all diagnosed in secondary care, as the basis for nested case-control samples.

8_ Knottnerus JA, Leffers JP.
The influence of referral patterns on the characteristics of diagnostic tests.
J Clin Epidemiol 1992;45:1143-54.

9_ Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Meulen van der JH et al.
Empirical evidence of design-related bias in studies of diagnostic tests.
JAMA 1999;282:1061-6.

10_ Reid MC, Lachs MS, Feinstein AR.
Use of methodological standards in diagnostic test research. Getting better but still not good.
JAMA 1995;274:645-51.

11_ Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J.
Sources of variation and bias in studies of diagnostic accuracy: a systematic review.
Ann Intern Med 2004;140:189-202.

12_ Begg CB, Greenes RA.
Assessment of diagnostic tests when disease verification is subject to selection bias.
Biometrics 1983;39:297-15.

13_ Mantel N.
Synthetic retrospective studies and related topics.
Biometrics 1973;29:479-86.

14_ Rothman KJ, Greenland S.
Modern epidemiology.
Philadelphia: Lincot-Raven Publishers, 1998.

15_ Oudega R, Moons KG, Hoes AW.
Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care.
Fam Pract 2005.

Patients & Methods

Description of the empirical data set

Data were derived from a large cross-sectional cohort study among adult patients suspected of deep vein thrombosis (DVT) in primary care. Details on the setting and data collection have been described previously[15]. In brief, the full cohort included 1295 consecutive patients who visited one of the participating primary care physicians because of symptoms compatible with DVT. Clinical suspicion of DVT was primarily based on the presence of a painful and swollen leg that existed no longer than 30 days. Patients were excluded from the study if pulmonary embolism was suspected.

The general practitioner systematically documented information on the patient's history and physical examination. Patient history included age, gender, history of malignancy, immobilisation and recent surgery. Physical examination included distension of collateral veins, swelling of the affected limb and difference in circumference of the calves (calculated as circumference of affected limb minus circumference of unaffected limb, further referred to as calf difference test). After the standardised history and physical examination, all patients were referred to one of three adherent hospitals to undergo D-dimer testing. In line with available guidelines and previous studies, the D-dimer test result was considered abnormal if the assay yielded a D-dimer level ≥ 500 ng/ml[16;17].

Finally, they all underwent the reference test, which was repeated compression ultrasonography (CUS) of the lower extremities. In patients with a normal first CUS measurement, the CUS was repeated after seven days. DVT was considered present if one CUS measurement was abnormal, which occurred in 289 patients. The echographist was blinded to the results of the patient history and physical examination.

Nested case-control samples

Nested case-control samples were drawn from the full cohort (n= 1295). In all samples all 289 cases with DVT were included. Controls were randomly sampled from the 1006 subjects without DVT. Four different case-control ratios were used, i.e. 1 control for each case (1:1), 2 controls for each case (1:2), 3 controls for each case (1:3) and 4 controls for each case (1:4). Hence, a sample with a case-control ratio of 1:1 contained 289 random subjects out of 1006 controls (sample fraction 1006/289 = 3.48) and 289 cases (in total 578 subjects). The sampling procedure was repeated 100 times for each case-control ratio.

Data analysis

To empirically document the validity of the nested case-control approach, we focussed on two important diagnostic tests for DVT, i.e. the dichotomous D-dimer test and continuous calf difference test. Measures of diagnostic accuracy with corresponding 95% confidence intervals of both tests were estimated for the four nested case-control ratios and compared with those obtained from the full cohort. The accuracy measures obtained from the nested case-control samples are shown in boxplots. Measures of diagnostic accuracy included positive and negative predictive value, sensitivity and specificity, positive and negative likelihood ratios and the odds ratio (OR) for the D-dimer test and the OR and the ROC area for the calf difference test. The ROC area can be interpreted as the probability that a randomly chosen patient with the outcome (DVT) will have a larger calf difference than a randomly chosen patient without DVT[18]. A ROC area of 0.5 indicates no

16_ Perrier A, Desmarais S, Miron MJ, de Moerloose P, Lepage R, Slosman D et al. Non-invasive diagnosis of venous thromboembolism in outpatients. Lancet 1999;353:190-5.

17_ Schutgens RE, Ackermark P, Haas FJ, Nieuwenhuis HK, Peltenburg HG, Pijlman AH et al. Combination of a normal D-dimer concentration and a non-high pretest clinical probability score is a safe strategy to exclude deep venous thrombosis. Circulation 2003;107:593-7.

18_ Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

19_ Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C et al. A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process. J Intern Med 1998;243:15-23.

20_ Kearon C, Ginsberg JS, Douketis J, Crowther M, Brill-Edwards P, Weitz JI et al. Management of suspected deep venous thrombosis in outpatients by using clinical assessment and D-dimer testing. Ann Intern Med 2001;135: 108-11.

21_ Swan SH, Shaw GM, Schulman J. Reporting and selection bias in case-control studies of congenital malformations. Epidemiology 1992;3:356-63.

22_ Hak E, Wei F, Grobbee DE, Nichol KL.
A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis.
J Clin Epidemiol 2004;57: 875-80.

23_ Essebag V, Genest J, Suissa S, Pilote L.
The nested case-control study in cardiology.
Am Heart J 2003;146:581-90.

24_ Lagnaoui R, Begaud B, Moore N, Chaslerie A, Fourrier A, Letenneur L et al.
Benzodiazepine use and risk of dementia: a nested case-control study.
J Clin Epidemiol 2002;55: 314-8.

25_ Meijer WE, Heerdink ER, Nolen WA, Herings RM, Leufkens HG, Egberts AC.
Association of risk of abnormal bleeding with degree of serotonin reuptake inhibition by antidepressants.
Arch Intern Med 2004;164: 2367-70.

discrimination, whereas an estimate of 1.0 indicates perfect discrimination. In the analysis of the nested case-control samples, controls were weighed by the sampling fraction corresponding to the case-control ratio (1:1 = 3.48; 1:2 = 1.74; 1:3 = 1.16; 1:4 = 0.87). For each case-control ratio, the 95% confidence intervals of diagnostic accuracy measures were obtained from the empirical distributions across the 100 samples drawn from the full cohort.

Analyses were performed using SPSS version 12.0 and S-plus version 6.0 software.

Results

Patient characteristics

The results of the D-dimer test and calf difference test in the full cohort and in the nested case-control samples are shown in Table 1. DVT was present in 289 (22%) patients. In the full cohort (n= 1295), the D-dimer test was abnormal in 892 (69%) patients and the mean difference in calf circumference was 2.3 cm. As expected, the proportions of an abnormal D-dimer test and mean calf difference were higher in the nested case-control samples than in the full cohort, since these variables are known predictors of DVT presence[15;19;20]. The differences decreased with an increasing proportion of controls in the nested case-control samples.

Measures of diagnostic accuracy

Table 2 shows the diagnostic accuracy measures and corresponding 95% confidence intervals of the D-dimer and calf difference tests as estimated from the full cohort. Sensitivity and negative predictive value were high for the D-dimer test, 0.94 and 0.96, respectively. Specificity (0.38) and positive predictive value (0.30) were low. The ROC area was 0.69 for the calf difference test and the OR was 1.44.

The median of the 100 estimates of diagnostic accuracy as estimated for the four nested case-control ratios were very similar to the corresponding estimates of the full cohort (Figure 2). For example, the negative predictive value of the D-dimer test was approximately 0.96 in the full cohort and the median of all four nested case-control ratios was also 0.96. The OR of the calf difference test was approximately 1.44 in the full cohort and the median of the nested case-control samples was also 1.44.

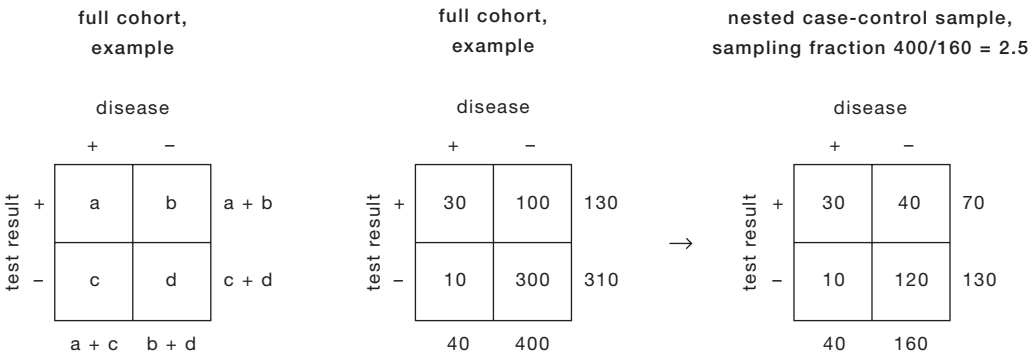


Figure 1. Example of drawing a nested case-control sample from a full cohort with known size. The case-control ratio was 1:4, 40 cases, 160 controls (sampling fraction (SF) = 400/160 =2.5). One can obtain valid diagnostic accuracy measures from the nested case-control sample, by multiplying the controls (b and d) with the sampling fraction. The positive predictive value (PPV) of a full cohort can be calculated with a / (a + b), in this example: 30 / (30 + 100) = 0.23. In the nested case-control sample the PPV can be calculated with a / (a + SF*b), in this example: 30 / (30 + 2.5*40) = 0.23. In a conventional case-control sample however, the controls are sampled from a source population with unknown size. Therefore, the sample fraction is unknown and a valid estimate of the PPV cannot be calculated. In this example, the PPV of a conventional case-control sample would be calculated with 30 / (30 + 40) = 0.43.

Table 1. Distribution of the D-dimer and calf difference test results in the full cohort and the nested case-control samples with various case-control ratios.Values represent absolute patient numbers (%) unless stated otherwise.

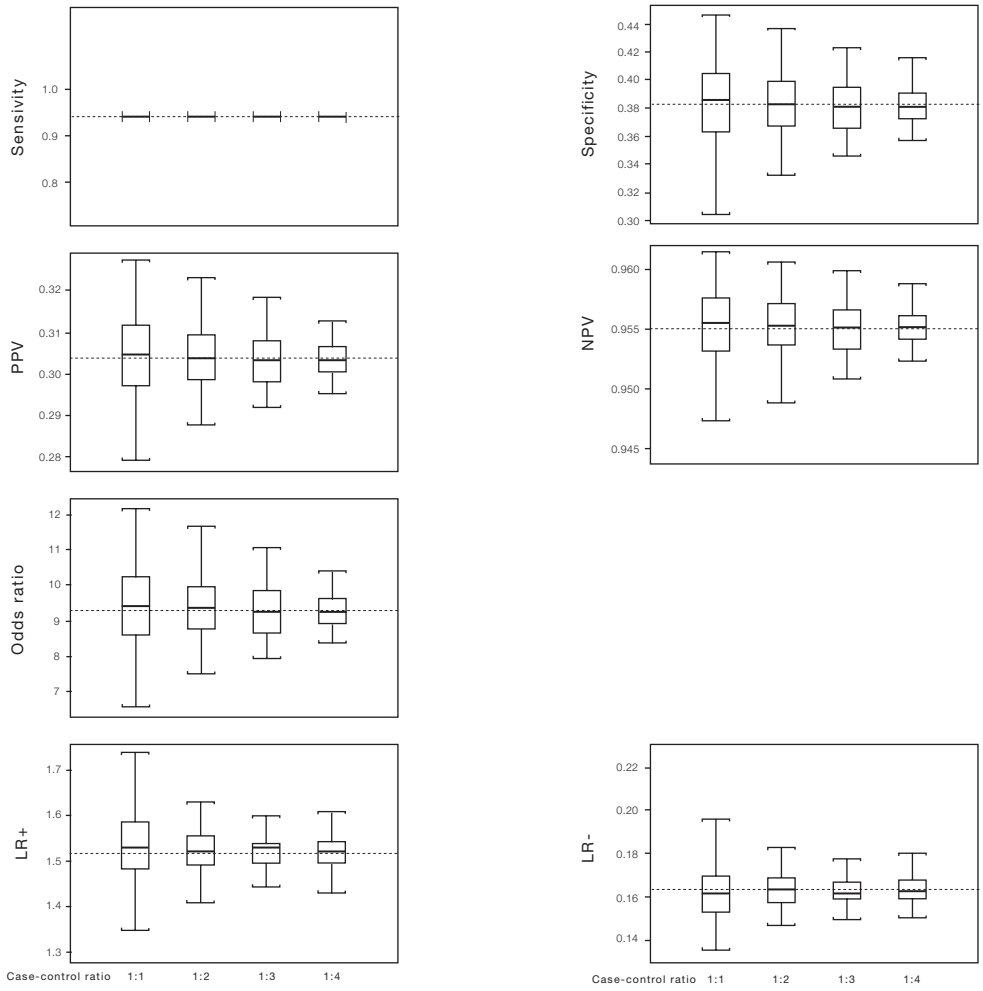
Characteristics	Full cohort	nested case-control samples with corresponding case-control ratios			
		1:1	1:2	1:3	1:4
	n=1295	n= 578	n= 867	n= 1156	n= 1445
D-dimer test abnormal	892 (69)	450 (78)	630 (73)	805 (70)	986 (68)
Calf difference (cm)*	2.3 (1.7)	2.6 (1.8)	2.5 (1.7)	2.3 (1.7)	2.3 (1.7)
Age (years)*	60.0 (17.6)	60.6 (17.3)	60.2 (17.5)	60.0 (17.5)	60.0 (17.6)
Male gender	467 (36)	232 (40)	325 (38)	421 (37)	516 (36)
Prevalence of DVT	289 (22)	289 (50)	289 (33)	289 (25)	289 (20)

*mean (standard deviation)

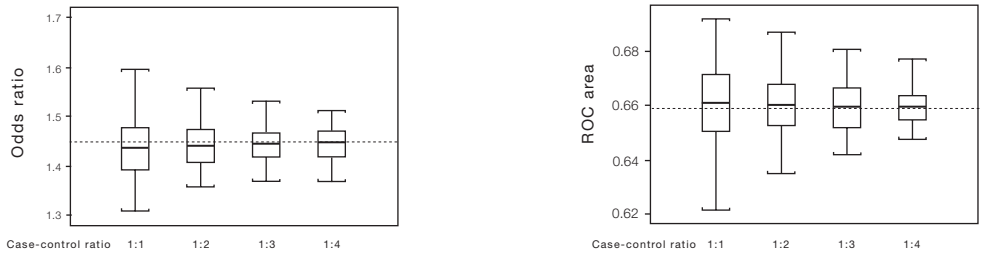
Table 2. Diagnostic accuracy measures with corresponding 95% confidence intervals for abnormal D-dimer and calf difference test results estimated in the full cohort.

Measures of diagnostic accuracy	D-dimer test (dichotomous)	Calf difference test (continuous per cm)
Sensitivity	0.94 (0.93-0.95)	–
Specificity	0.38 (0.37-0.40)	–
PPV	0.30 (0.28-0.33)	–
NPV	0.96 (0.90-1.0)	–
LR+	1.52 (1.52-1.52)	–
LR-	0.16 (0.15-0.18)	–
Odds ratio	9.33 (5.70-15.3)	1.44 (1.33-1.56)
ROC area	–	0.69 (0.65-0.72)

- = not applicable
PPV = positive predictive value
NPV = negative predictive value
LR+ = likelihood ratio of a positive test result
LR- =likelihood ratio of a negative test result
ROC area = area under the receiver operating characteristic curve



a. D-dimer test



b. Calf difference test

Figure 2. Graphical presentation of diagnostic accuracy measures of a) the D-dimer test and b) calf difference test for the 100 nested case-control samples with case-control ratios ranging from 1:1 to 1:4. The boxes indicate median values and corresponding interquartile ranges (25th and 75th percentile). Whiskers indicate 95% confidence intervals. The dotted lines represent the values estimated from the full cohort.

Concluding remarks

Results from conventional diagnostic case-control studies have been questioned because of methodological limitations[1;3;8-10;12]. Consequently, this approach has not been recommended in the recent STARD guideline[5;6]. We argue that those limitations do not apply to nested case-control studies. Rather, nested case-control studies offer a valid and efficient alternative to diagnostic studies in, cross-sections of, cohort studies. Importantly, case-control studies nested in an existing cohort with known sample size cohort -and in which the 'true' disease status is determined by a reference test- are not subject to verification bias, and can yield absolute disease probabilities[1;3;8;9;12;21]. This was confirmed by our analysis of a diagnostic study in patients suspected of deep venous thrombosis in which diagnostic accuracy measures of the full cohort and nested case-control samples were very similar. Expectedly, the width of the 95% confidence intervals decreased with increasing number of controls, making the measures estimated in the larger case-control samples more precise. The findings support the view that the nested case-control approach offers an attractive and more efficient design for diagnostic studies and should be reappraised in current method guidelines.

Our findings are in agreement with other studies. Several recent etiologic papers demonstrated that results from nested case-control studies are virtually identical to results based on a full cohort analysis, if appropriate statistical methods are applied[22-25].

In conclusion, a nested case-control design reduces the number of unnecessary measurements as compared to the full cohort approach with equivalent validity. We showed that the nested case-control design offers investigators a valid and efficient alternative for a full cohort approach in diagnostic research. This may be particularly important when the results of the test under study are costly or difficult to collect.

chapter

5

Validating and updating a prediction rule for neurological sequelae after childhood bacterial meningitis

Introduction

In spite of optimal intensive care facilities and antibiotic treatment, childhood bacterial meningitis still causes persistent neurological sequelae in 10-20% and mortality in about 5% of the patients[1-4]. Early prediction of an adverse outcome may help to determine which children with bacterial meningitis are at high risk for developing such outcomes and need therefore more intensive or longer follow-up.

Recently, we developed a prediction rule to predict neurological sequelae or death within six months after childhood bacterial meningitis[5]. This prediction rule included four predictors which were selected from a larger number of candidate predictors (13). The rule was developed with data of 93 patients with bacterial meningitis from two pediatric teaching hospitals in The Netherlands (derivation set). The rule showed good performance reflected in good calibration (i.e. good agreement between predicted risks of neurological sequelae or death and observed frequencies of this outcome) and good discrimination (i.e. the ability to distinguish patients with neurological sequelae or death from healthy survivors). Prediction rules generally tend to perform better on patients from which the rule has been derived than on new patients. Hence, before a prediction rule can be applied in practice, it must be tested in new patients that represent the same domain, e.g. type of patients for which the rule is developed (external validation)[6-9]. This particularly applies when the number of patients in the derivation data set is small and relatively many candidate predictors are considered[10-14], which was the case with the present prediction rule.

The aim of the present study was therefore 1) to test the generalisability of the prediction rule in a large sample of children with bacterial meningitis selected from almost all hospitals in The Netherlands (validation set), and 2) to update the prediction rule using the combined derivation and validation sets.

Patients and Methods

Derivation study

The prediction rule was developed with 93 patients, aged between 1 month and 15 years. They presented with meningeal signs in two teaching hospitals in The Netherlands between 1988 and 1998, and they had a final diagnosis of bacterial meningitis[5;15]. The study population, the applied methods, the chosen outcomes and the main results have been published[5;15]. In brief, 247 patients who had bacterial meningitis were initially selected. Meningitis cases caused by *Haemophilus Influenzae type B* were excluded since this type has virtually disappeared, due to routine vaccination. Patients with pre-existent neurological diseases were also excluded, as it would not be possible to determine afterwards the cause of the neurological sequelae. The status of the remaining 170 children was investigated approximately seven months (range 0.02-3.3 years) after the onset of bacterial meningitis, to determine whether they had developed neurological sequelae or died5. There were 23 events; twenty one patients developed neurological sequelae and two patients died. The prediction rule for neurological sequelae was developed with these 23 cases and a random sample of 70 controls from the remaining 147 healthy recoveries. Accordingly, the derivation study used a case-control design nested in the cohort of 170 children with bacterial meningitis.
Starting with 13 candidate predictors selected from the literature, stepwise multivariate regression analysis yielded a final prediction rule with four predictors.

1_ Baraff LJ, Lee S, Schriger D. Outcomes of bacterial meningitis in children: a meta-analysis. Pediatr Infect Dis J 1993;12:389-94.

2_ Oostenbrink R, Maas M, Moons KG, Moll HA. Sequelae after bacterial menigitis in childhood. Scan J Infect Dis 2002;34: 379-82.

3_ Grimwood K, Anderson P, Anderson V, Tan L, Nolan T. Twelve year outcomes following bacterial meningitis: further evidence for persisting effects. Arch Dis Child 2000;83:111-6.

4_ Bedford H, de Louvois J, Halket S, Peckham C, Hurley R, Harvey D. Meningitis in infancy in England and Wales: follow up at age 5 years. BMJ 2001;323:533-6.

5_ Oostenbrink R, Moons KG, Derksen-Lubsen G, Grobbee DE, Moll HA. Early prediction of neurological sequelae or death after bacterial meningitis. Acta Paediatr 2002;91:391-8.

6_ Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130: 515-24.

7_ Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453-73.

8_ Houwelingen van JC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19:3401-15.

9_ Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995;14:1999-2008.

Predictors were male gender, presence of atypical convulsions in medical history (defined as seizure with either duration longer than 15 minutes, non-generalised jerks or multiple seizures within 24 hours), high body temperature at physical examination, and pathogen of infection (*N. Meningitidis* or *S. Pneumoniae*). The formula of the prediction rule after bootstrapping was[5]:

*Probability of neurological sequelae = 1/(1 + exp [-(25.2 + 1.48*male gender + 2.27*presence of atypical convulsions – 0.75*body temperature at physical examination + 3.12*S. pneumoniae + 1.48*N. meningitidis)])* **(1)**

In this, 25.2 represents the rule's intercept whereas the other numbers are the regression coefficients (weights) of each corresponding predictor; the exponent of a regression coefficient equals the odds ratio. To estimate the risk for an individual patient, the patient value of each predictor is multiplied by its regression coefficient. For example, a boy with bacterial meningitis caused by *N. meningitidis*, with absence of atypical convulsions in patient history and a body temperature at physical examination of 38.5˚C has a risk of developing neurological sequelae within 3 years of 1/(1 + exp [-(25.2 + 1.48 – 0.75*38.5 + 1.48)]) = 0.33. The ROC area was 0.87 (95%CI 0.78-0.96) after correction for overoptimism with bootstrapping. The general agreement between observed and predicted risks was satisfactory as visualised by the rule's calibration plot5. This was confirmed by a non-significant Hosmer-Lemeshow test for goodness-of-fit (p-value = 0.58)[16].

Validation data

Data for the present validation study on which the prediction rule (formula 1) was tested (validation set), were retrieved from a large retrospective cohort study that aimed to assess academic and behavioral limitations at school age in survivors of bacterial meningitis. For details on the study population, the applied methods and questionnaires, the chosen outcome definitions, and the main results we refer to the literature [17-19]. In brief, patient data were selected from files of the Netherlands Reference Laboratory for Bacterial Meningitis. All children who were born between January 1986 and December 1994 and who recovered from bacterial meningitis between January 1990 and December 1995 were selected. Bacterial meningitis was caused by either *N. meningitidis, S. pneumoniae, S. agalacticae, E. coli or L. monocytogenes*. Patients with meningitis caused by *H. influenzae type B* and other less common pathogens were excluded. Patients with meningitis secondary to immunodeficiency states, central nervous system surgery, cranial trauma, cerebrospinal fluid shunt infection or relapsing meningitis were also excluded. Other exclusion criteria were the presence of cognitive or behavioral problems prior to meningitis and diseases developed after meningitis (e.g. cancer), which could have caused cognitive or behavioral problems. In total, 628 patients with bacterial meningitis were included in the original cohort and thus were available for inclusion in the present validation study[17;18].

Since the validation cohort included only survivors of bacterial meningitis, the outcome for validation of the prediction rule was limited to neurological sequelae. Patients were considered to have neurological sequelae if they met one or more of the following criteria: 1) sensorineural hearing impairment (≥ 25dB); 2) epilepsy/mental retardation (IQ <80 and a professional's diagnosis of mental retardation or hemi-/di-/tetraparesis; 3) cortical visual defect or hydrocephalus. For each patient the outcome was determined based on information from the

10_ Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE et al.
External validation is necessary in prediction research: A clinical example.
J Clin Epidemiol 2003;56: 826-32.

11_ Harrell FE, Lee KL, Mark DB.
Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros.
Stat Med 1996;15:361-87.

12_ Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE.
Redundancy of single diagnostic test evaluation.
Epidemiology 1999;10:276-81.

13_ Steyerberg EW, Eijkemans MJ, Habbema JD.
Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis.
J Clin Epidemiol 1999;52: 935-42.

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

15_ Oostenbrink R, Moons KG, Donders AR, Grobbee DE, Moll HA.
Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures.
Acta Paediatr 2001;90:611-7.

16_ Hosmer D, Lemeshow S.
Applied logistic regression.
New York: John Wiley & Sons, Inc., 1989.

medical records and questionnaires completed by the patients' parents. On average, the time from onset of bacterial meningitis until assessment of neurological sequelae was 6.2 years (range 3-10 years).

External validation of the prediction rule

To study external validation of the prediction rule (formula 1), calibration and discrimination were quantified in the validation set. Calibration indicates to what extent the observed frequencies of neurological sequelae agree with risks for neurological sequelae as predicted by the rule. A graphical impression of calibration was obtained by plotting the observed frequencies versus the predicted risks (calibration plot). In addition, calibration was statistically tested across deciles of predicted risks with the Hosmer-Lemeshow goodness-of-fit test[16]. Discrimination was studied with the ROC area, which reflects the ability of the prediction rule to discriminate patients with neurological sequelae from those without. It ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination).

Updating the prediction rule

In order to improve the performance, we updated the prediction rule after combining the derivation set (n=93) and the validation set (n=628). We deleted the two dead patients from the derivation set, because the validation data set also included only survivors of meningitis. Furthermore, three other patients (without neurological sequelae) were excluded from the derivation set, since they were also included in the validation set. Hence, the total data set included 716 subjects of which 87 had developed neurological sequelae.

We selected eight predictors, based on literature and clinical experience[2;20-23]. These included male gender, age at infection, atypical convulsions in medical history, presence of petechiae and/or ecchymoses at physical examination, body temperature at physical examination, type of pathogen *(N. meningitidis, S. pneumoniae*, or *other*), use of anti-epileptica for more than 2 days during hospital admission for bacterial meningitis (as a proxy for presence of persistent convulsions) and use of mechanical ventilation during hospital admission.
The continuous predictors age and body temperature were plotted against the risk of neurological sequelae (on the log odds scale) to study linearity of these predictors.

All candidate predictors were simultaneously entered into a multivariate logistic regression model. Variables with a weak predictive value were deleted from the model with stepwise backward selection. This results in a model with the strongest predictors of neurological sequelae. A predictor was excluded from the model if the likelihood ratio test showed a p-value > 0.15[11]. Bootstrapping techniques have been advocated to adjust the model's estimated regression coefficients and predictive performance for overoptimism[24-27]. We repeated the modeling process (including the stepwise predictor selection) in 100 bootstrap samples. This yielded a shrinkage factor for the regression coefficients and an optimism corrected estimate of the ROC area. Eventually, the prediction rule was transformed into a nomogram to facilitate the computation of the predicted risk of developing neurological sequelae for a child with bacterial meningitis.

In the combined data set some investigated predictors (not the outcome variable) had missing values. It has been shown that deleting subjects with a missing value on one of the predictors (complete case analysis) leads to bias and loss of power. Therefore, it is better to impute the missing values[11;28]. Accordingly, missing values were imputed using the regression method available in SPSS for windows.

17_ Koomen I, Grobbee DE, Jennekens-Schinkel A, Roord JJ, van Furth AM. Parental perception of educational, behavioural and general health problems in school-age survivors of bacterial meningitis. Acta Paediatr 2003;92:177-85.

18_ Koomen I, Grobbee DE, Roord JJ, Donders R, Jennekens-Schinkel A, van Furth AM. Hearing loss at school age in survivors of bacterial meningitis: assessment, incidence, and prediction. Pediatrics 2003;112:1049-53.

19_ Koomen I, Grobbee DE, Roord JJ, Jennekens-Schinkel A, van der Lei HDW, Kraak MA et al. Prediction of academic and behavioural limitations in school-age survivors survivors of bacterial meningitis. Acta Paediatr 2004.

20_ Kaaresen PI, Flaegstad T. Prognostic factors in childhood bacterial meningitis. Acta Paediatr 1995;84:873-8.

21_ Kornelisse RF, Westerbeek CM, Spoor AB, van der HB, Spanjaard L, Neijens HJ et al. Pneumococcal meningitis in children: prognostic indicators and outcome. Clin Infect Dis 1995;21:1390-7.

22_ Pikis A, Kavaliotis J, Tsikoulas J, Andrianopoulos P, Venzon D, Manios S. Long-term sequelae of pneumococcal meningitis in children. Clin Pediatr (Phila) 1996;35: 72-8.

23_ Pomeroy SL, Holmes SJ, Dodge PR, Feigin RD. Seizures and other neurologic sequelae of bacterial meningitis in children. N Engl J Med 1990;323:1651-7.

Table 1. Distribution of patient characteristics in the derivation and validation set.

Characteristic	Derivation study		
	Nested case-control (n=93)	Full cohort (n=170)*	Validation set (n=628)
Predictors of the original rule			
Male gender	48 (52)	88 (52)	356 (57)
Atypical convulsions in patient history	12 (13)	22 (13)	2 (0.3)
Presence of petechiae/ecchymoses	40 (43)	73 (43)	338 (54)
Body temperature at physical examination (°C)†	39.0 (1.1)	39.0 (1.1)	39.1 (1.0)
Pathogen type			
S. pneumonia	16 (17)	29 (17)	103 (16)
N. meningitidis	64 (69)	117 (69)	495 (79)
Other	13 (14)	24 (14)	30 (5)
Other patient characteristics			
Age at infection (years) ‡	2.8 (0.9-5.8)	3.1 (0.8-6.2)	1.9 (0-9.5)
Use of anti-epileptica > 2 days during hospital admission for bacterial meningitis	8 (9)	16 (9)	61 (10)
Mechanical ventilation during hospital admission	11 (12)	20 (12)	39 (6)
Outcome neurological sequelae			
Total	23 (25)	23 (14)	66 (11)
Hearing loss or deafness	14 (15)	14 (8)	43 (7)
Epilepsy, mild motor deficits, hemi/ di/ tetraparesis or mild mental retardation	3 (3)	3 (2)	15 (2)
(Severe) mental retardation and tetraplegia	4 (4)	4 (2)	10 (2)
Cortical visual defect, hydrocephalus or death	2 (2)	2 (1)	10 (2)

Values represent absolute patient numbers (%) unless stated otherwise.
* The analysis of the derivation study was based on a nested case-control design; subjects of the derivation set (n=93; 23 cases, 70 controls, sample ratio 1:3) were nested in a cohort study (n=170; 23 cases, 147 controls). The 70 sampled controls were weighted by a factor 2.1 yielding a virtual full cohort, which was needed to reflect the data distribution in the original full cohort. This facilitated valid comparison of the derivation and validation cohorts.
† Mean (standard deviation)
‡ Median (range)

Table 2. Distribution of candidate predictor values across cases and non-cases with univariate associations of each predictor with neurological sequelae in the total data set (n=716).

Candidate predictor	Sequelae present (n= 87)	Sequelae absent (n= 629)	Univariate odds ratio (95%CI)
Male gender‡	56 (64)	344 (55)	1.5 (0.9-2.4)
Age at infection (years)‡	1.8 (2.3)†	2.7 (2.4)†	0.82 (0.72-0.92)
Atypical convulsions	6 (7)	7 (1)	4.5 (1.8-12)
Petechiae/ ecchymoses	17 (20)	360 (57)	0.33 (0.21-0.54)
Body temperature at physical examination < 40 ºC	38.3 (1.1)†	38.7 (0.8)†	0.62 (0.48-0.81)
Body temperature at physical) examination ≥ 40 ºC	40.5 (0.40)†	40.3 (0.35)†	3.3 (0.97-11.1)
Pathogen type			
N. meningitidis	40 (46)	512 (81)	0.20 (0.12-0.30)
S. pneumoniae	40 (46)	77 (12)	6.1 (3.4-4.9)
Other	7 (8)	40 (6)	*
Use of anti-epileptica >2 days	26 (30)	40 (7)	5.9 (3.4-10.2)
Mechanical ventilation during hospital admission	10 (11)	37 (6)	2.1 (1.0-4.3)

Values presented in the columns 'Sequelae present' and 'Sequelae absent' are absolute patient numbers (%), unless stated otherwise
†Mean (standard deviation)
‡not selected in multivariable analysis (p-value > 0.15)
*reference category

Table 3. Multivariate association of each predictor with neurological sequelae after bootstrapping in the derivation and combined data set. Values represent odds ratios with accompanying 95% confidence intervals.

Predictor	Combined data set (n=716)	Derivation set (n=93)
Male gender	*	4.4 (0.9-21)
Atypical convulsions	4.1 (1.1-16)	9.7 (1.0-99)
Petechiae/ ecchymoses	0.31 (0.2-0.50)	*
Body temperature at physical examination (per ºC)	0.58 (0.4-0.76)† 2.9 (0.9-8.7)‡	0.5 (0.2-0.9)
Pathogen type		
N. meningitidis	1.5 (0.5-4.2)	4.4 (0.5-42)
S. pneumoniae	*4.1 (1.5-11)*	*22.6 (1.3-394)*
Other	#	#
Use of anti-epileptica >2 days	2.1 (1.2-4.4)	*

* not selected in multivariable analysis (p-value > 0.15)
† for body temperature < 40 ºC
‡ for body temperature ≥ 40 ºC
# reference category

All calculations were performed using SPSS for Windows, version 12.0 and Splus version 6.1 software.

Results

Table 1 shows the distribution of patient characteristics, including the predictors of the prediction rule and the outcome, for the derivation and validation set. As the derivation study included a case-control analysis nested in a cohort, we presented the distribution of the patient characteristics for the case-control sample, i.e. the derivation set (n=93), and for the (virtual) full cohort (n=170). The derivation and validation sets were rather similar, although the validation set (n= 628) contained slightly younger patients, with only two patients with a history of atypical convulsions (less than 1% compared to 13% in the virtual derivation cohort) and more often *N. meningitidis* as the causative pathogen. Furthermore, neurological sequelae occurred in 14% of the patients in the cohort from which the derivation set was derived compared to 11% in the validation set.

External validation of the prediction rule

The calibration of the prediction rule was very poor in the validation set, as is shown in Figure 1. The ideal plot shows the situation when the observed frequencies and predicted risks are in perfect agreement. Predicted risks between 0 and 40% had all observed frequencies of around 10%. The rule clearly yielded too extreme predictions. This situation is typical when regression coefficients of the predictors are overfitted. The Hosmer-Lemeshow test statistic was statistically significant (p-value < 0.001), which also indicates a poor calibration in the validation set. The ROC area was 0.65 (95%CI: 0.57-0.72). This was statistically significant lower compared to the ROC area in the derivation set (0.87, 95%CI: 0.78-0.96, p-value < 0.01)[29].

Updating the original prediction rule using the combined data sets

Table 2 shows the distribution of all predictors across patients with neurological sequelae (n=87) and without (n=629) in the combined data sets (n=716). Also, the univariate associations are shown. All candidate predictors were associated with the occurrence of neurological sequelae. The continuous predictor age showed a linear association when plotted against the risk of neurological sequelae and was entered as a linear term in the multivariate analysis. Body temperature was entered as two piecewise linear terms in multivariate analysis; one for body temperature < 40 °C and one for body temperature ≥ 40 °C.

The stepwise backwards selection resulted in a model with five of the eight candidate predictors (Table 3, last column). These were atypical convulsions in medical history, body temperature at physical examination, presence of petechiae and/or ecchymoses, type of pathogen (*N. meningitidis, S. pneumoniae* or *other*), use of anti-epileptica for more than 2 days during hospital admission for bacterial meningitis as a proxy for persistent convulsions, and use of mechanical ventilation during hospital admission. The bootstrap procedure yielded a shrinkage factor for the regression coefficients of 0.90, which is acceptable and rather close to 1. The ROC area of the updated rule was 0.79 (95%CI: 0.73-0.84) before bootstrapping and 0.77 (0.72-0.82) afterwards.

24_ Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983;78: 316-31.

25_ Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall, 1993.

26_ Harrell FE. Regression modeling strategies. New York: Springer-Verlag, 2001.

27_ Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54: 774-81.

28_ Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142: 1255-64.

29_ 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.

30_ Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. Cancer Treat Rep 1985;69: 1071-7.

31_ Flack VF, Chang PC. Frequeny of selecting noise variables in subset regression: a simulation study. The American Statistician 1987;41:84-6.

The formula of the updated rule after bootstrapping was:

*Probability of neurological sequelae = 1/(1+exp -(-2.72 + 1.42*presence of atypical convulsions - 1.18*presence of petechiae/ecchymoses – 0.54*body temperature at physical examination < 40 ºC + 1.05*body temperature at physical examination ≥ 40 ºC + 0.39*N. meningitidis + 1.42*S. pneumoniae + 0.84*use of anti-epileptica >2 days))* **(2)**

Similar to formula 1, -2.72 is the intercept of the rule (estimated after weighing the controls in the original derivation set with factor 2.1) and the other numbers are the regression coefficients. One can estimate the risk that a child will develop neurological sequelae, using formula 2 or with the nomogram of the updated rule (Figure 2). The legend of Figure 2 shows how to use the nomogram for individual patients in practice.

Discussion

We validated a previously developed prediction rule for neurological sequelae after childhood bacterial meningitis, in a much larger cohort (validation set) than the cohort in which the rule was developed (derivation set). The prediction rule yielded very poor calibration and discriminative performance (ROC area= 0.65) in the validation set. The prediction rule appeared to be overfitted and much too optimistic and thus of limited use for general pediatric care. Therefore, we updated the original rule using the combined data of the derivation and validation sets. The updated rule contained two additional predictors, notably use of anti-epileptica > 2 days during hospital admission and presence of petechiae and/or ecchymoses. Gender was no longer included. The updated rule had much better discriminative performance (ROC area= 0.77).

Various reasons may explain the decrease in performance of the prediction rule that was observed in the validation set. The most likely reason is that the rule in the derivation study was so-called overfitted. In principle, no more than one variable per ten events (1 to 10 rule) should be considered when developing a prediction model[30-34]. This means that no more than 23 events/10 = 2 to 3 predictors should have been considered in the derivation set, whereas 13 candidate predictors were considered. Accordingly, the chance of finding spurious associations between the investigated predictors and the outcome in the derivation study was high. This has likely resulted in overfitted regression coefficients and too optimistic measures of predictive performance[10;13;30;35]. Although in the derivation study attempts were made to adjust for overfitting by applying bootstrapping techniques, this apparently did not result in a stable and generalisable prediction rule[6;32].

A second reason is that the validation set comprised patients of neurological sequelae only, whereas the derivation set also included death as an outcome. However, there were only two fatal cases in the derivation set. It is, in our view, unlikely that these differences in outcome assessment could explain the substantial decrease in performance of the original rule as observed in the large validation set.

A third reason for the poor performance of the prediction rule in the validation set could be the difference in the moment of outcome assessment and other casemix differences between the derivation and validation set. With respect to the outcome assessment, the children in the derivation set were assessed at a median of seven months post-meningitis whereas children of the validation set

32_ Simon R, Altman DG.
Statistical aspects of prognostic factor studies in oncology.
Br J Cancer 1994;69:979-85.

33_ Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR.
A simulation study of the number of events per variable in logistic regression analysis.
J Clin Epidemiol 1996;49: 1373-9.

34_ Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD.
Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets.
Stat Med 2000;19:1059-79.

35_ Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD.
Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets.
Med Decis Making 2001;21:45-56.

36_ Grimwood K, Anderson VA, Bond L, Catroppa C, Hore RL, Keir EH et al.
Adverse outcomes of bacterial meningitis in school-age survivors.
Pediatrics 1995;95:646-56.

37_ Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG.
Internal and external validation of predictive models: a simulation study of bias and precision in small samples.
J Clin Epidemiol 2003;56: 441-7.

38_ Steyerberg EW, Borsboom GJ, Van Houwelingen HC, Eijkemans MJ, Habbema JD.
Validation and updating of predictive logistic regression models: a study on sample size and shrinkage.
Stat Med 2004;23:2567-86.

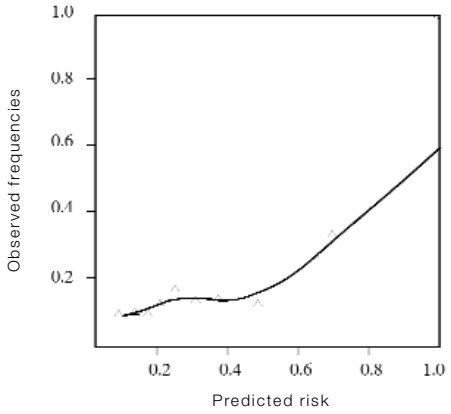


Figure 1. Calibration curve of the original prediction rule when tested on the validation set. Triangles indicate the observed frequencies of neurological sequelae per decile of predicted risks. The solid line shows the smoothed association between the predicted risks and observed frequencies. Ideally, this line equals the dashed line.

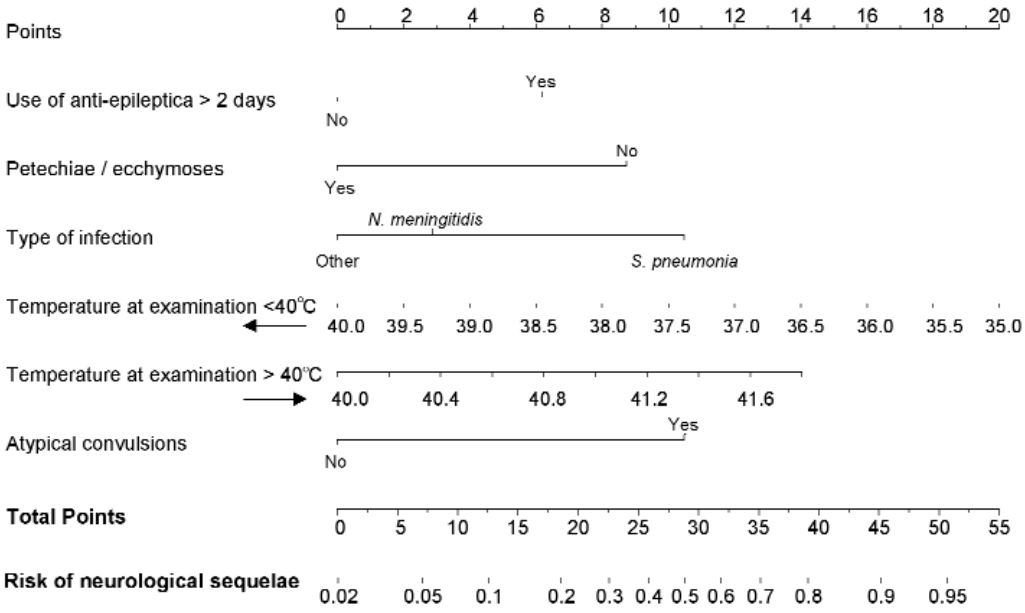


Figure 2. Nomogram relating the predictors of the combined model to the probability of developing neurological sequelae. As an example of using the nomogram, a child with use of anti-epileptica during hospital admission due to bacterial meningitis (which corresponds to 6 points as determined from the 'Points' scale on top of the figure), who had petechiae and/or ecchymoses (no points) who was infected with *S. pneumoniae* (11 points), who had a body temperature of 41.0 °C (8 points) and no atypical convulsions in medical history (no points) receives a 'Total Points' score of 6 + 11 + 8 = 25 points. Using the lower two scales of the nomogram, this score corresponds to a probability of developing neurological sequelae of about 38%.

were assessed at a median of six years post-meningitis. Furthermore, in the validation set, children were at school-going age at the time of outcome assessment whereas the children in the derivation set were assessed at a median age of 2.8 years. Accordingly, mild deficits, like mild mental retardation, may not yet have been detected in the derivation set and could have been developed at older age[36]. This could have resulted in a underestimation of the (incidence of) neurological deficits in the original derivation study. Regarding the other differences in case mix, in the derivation study bacterial meningitis patients were selected from two teaching hospitals and included only cases who primarily presented with 'meningeal signs', and not with other symptoms or signs such as convulsions[5;15]. However, we should note that both hospitals also function as secondary care hospitals: about half of their patients are referred and half are self-referrals[5;15]. The validation set included the survivors of all bacterial meningitis cases irrespective of the initial presentation, who were selected from all (110) hospitals in The Netherlands. Accordingly, the validation set included the cases from the entire disease spectrum and from secondary and tertiary centers.

The differences between the two data sets are apparent from Table 1. However, regardless the exact reason for the differences, the fact remains that the original prediction rule - as previously published - shows very poor accuracy and thus generalisability across other bacterial meningitis cases. This was the very reason to combine the derivation and validation set to update the original prediction rule based on a larger amount of data and to study whether other predictors of neurological sequelae could be identified[8;10;37;38]. The combined data set included a 7 times larger population of meningitis cases than the derivation set, and thus better reflects the heterogeneous distribution of survivors of bacterial meningitis cases in clinical practice. Accordingly, the updated rule developed from the combined data set is certainly less overfitted and very likely has improved generalisability compared to the original rule.

Nevertheless, before wide spread application in practice can be advocated, we do suggest that the updated rule should first be validated in other bacterial meningitis populations to test its generalisability, because the definition and use of the predictors in the updated rule may vary across institutions.

In the updated rule, atypical convulsions, body temperature and pathogen type were again selected from the combined data set, reassuring their relevance as predictors. However, as expected from the much larger number of outcome events (87 instead of 23), the magnitude of their associations had decreased. Their odds ratios were closer to 1, i.e. less extreme and thus less overfitted than in the derivation study. Male gender was not a predictor in the updated prediction rule, whereas anti-epileptica use for more than 2 days and presence of petechiae and/or ecchymoses were additional predictors (Table 3). The additional inclusion of these predictors in the updated rule and the exclusion of male gender may again be explained by the larger number of study subjects and the larger heterogeneity in case mix in the combined data set. Obviously, atypical convulsions and the use of anti-epileptics partly overlap. The fact that both predictors were significant in the multivariable model indicates that they both had independent predictive value for neurological sequelae, and thus should be included in the final prediction model.

Predictors of neurological sequelae after childhood bacterial meningitis have been debated extensively in the literature[2;5;18;20;21;23;36;39-44]. However, only a small amount of published work in the post *H. influenzae type B* era has been dedicated to predictors of these neurological sequelae[2;18;19]. Our results partly agree and partly disagree with the literature. Notably, the presence of petechiae

39_ Valmari P, Peltola H, Ruuskanen O, Korvenranta H. Childhood bacterial meningitis: initial symptoms and signs related to age, and reasons for consulting a physician. Eur J Pediatr 1987;146:515-8.

40_ Woolley AL, Kirk KA, Neumann AM, McWilliams SM, Murray J, Freind D et al. Risk factors for hearing loss from meningitis in children: the Children's Hospital experience. Arch Otolaryngol Head Neck Surg 1999;125:509-14.

41_ Bedford H. Prevention, treatment and outcomes of bacterial meningitis in childhood. Prof Nurse 2001;17:100-2.

42_ Dodge PR, Davis H, Feigin RD, Holmes SJ, Kaplan SL, Jubelirer DP et al. Prospective evaluation of hearing impairment as a sequela of acute bacterial meningitis. N Engl J Med 1984;311:869-74.

43_ Klinger G, Chin CN, Beyene J, Perlman M. Predicting the outcome of neonatal bacterial meningitis. Pediatrics 2000;106:477-82.

44_ Feigin RD, McCracken GH, Klein JO. Diagnosis and management of meningitis. Pediatr Infect Dis J 1992;11:785-814.

45_ Fortnum HM. Hearing impairment after bacterial meningitis: a review. Arch Dis Child 1992;67: 1128-33.

46_Grimwood K, Nolan TM, Bond L, Anderson VA, Catroppa C, Keir EH. Risk factors for adverse outcomes of bacterial meningitis. J Paediatr Child Health 1996;32:457-62.

and/or ecchymoses appeared to decrease the risk of developing neurological sequelae, given the odds ratio smaller than 1. The inclusion of this predictor in the updated rule is in agreement with a study on predicting hearing loss after childhood bacterial meningitis[18]. That study similarly found that *S. pneumoniae* was the most virulent pathogen for developing hearing loss in survivors of pneumococcal meningitis. This was not only confirmed in our study but also by many others[1;4;21;40;45]. In contrast, two previous studies[19;40] found male gender as a predictor for neurological sequelae. In our study male gender was as well a predictor but only in the univariable analysis. In multivariable analysis gender did not add predictive power beyond the other predictors. This was also the case though in two other studies[20;46]. As discussed above, differences between our results and previous studies may result from dissimilarities in study population due to e.g. differences in age at infection, in time of assessment of sequelae and in severity of studied sequelae.

Although our combined data set was 7 times larger than the derivation set, it still included only 87 events. To satisfy the above described 1 to 10 rule[30-34], we specifically selected only 8 predictors as based on the existing literature and our previous studies on the data[2;5;17;19] In these previous studies many more potential predictors of neurological sequelae, e.g. duration of symptoms prior to diagnosis, use of antibiotics, and blood parameters, were studied but appeared no important predictors. Given the 83 events in the combined data set, we did not study these other potential predictors again. However, they may still play some role in the prediction of neurological sequelae after childhood bacterial meningitis. Further research should indicate whether the newly developed rule can be further updated and improved by these additional predictors.

In conclusion, external validation of a previously developed prediction rule for neurological sequelae after childhood bacterial meningitis showed very poor performance when applied to a larger validation cohort of bacterial meningitis patients. The rule required updating. The updated prediction rule showed a significantly better performance. Before wide spread application in clinical practice, however, we suggest to perform some additional validation studies to test the generalisability of the updated rule across other bacterial meningitis populations. Our analyses again demonstrate the importance of using new (validation) data to test existing prediction rules[8;10;37;38]. Rather than viewing a validation data set as a separate study to estimate an existing rule's performance, validation data can often better be combined with data of previous derivation studies to generate more robust prediction models.

chapter
6

Genetic programming or multivariable logistic regression in diagnostic research

Introduction

In the past decade there has been an increased interest in medical prediction research to answer prognostic and diagnostic questions. Generally, such research aims to develop a so-called prediction rule to predict a particular outcome as accurate as possible, preferably with a minimum of information or predictors. In diagnostic prediction research the outcome includes the presence of a disease and in prognostic prediction research the future occurrence of a certain event. With the increasing availability of electronic patient records the interest in medical prediction research will further increase since electronic records facilitate the application of prediction rules in medical practice.

The most widely used method to develop prediction rules or models in clinical epidemiology is multivariable logistic regression analysis[1-7]. In the past decade, new methods such as classification and regression trees (CART) and neural networks have been introduced for this purpose. However, it has repeatedly been shown that both methods do not produce prediction rules that achieve higher predictive accuracy than rules developed by multivariable logistic regression[8-13]. Recently, the technique of genetic programming has emerged. Genetic programming is a search method inspired by the process of natural evolution and may be used to solve complex associations between large numbers of variables[14-16]. This feature makes genetic programming also suitable for prediction research.

Genetic programming is not restricted to any fixed model structure. Therefore, it may theoretically yield a model with higher predictive accuracy compared to a model obtained by conventional logistic regression analysis. However, the flexibility of a logistic model can also be increased by including cubic splines for continuous variables (rather than only the linear terms) and interaction terms, potentially enhancing the model's predictive accuracy[4;6;17]. However, this is not commonly done as it makes the model less easy to interpret.

Like neural networks, genetic programming originates from the field of artificial intelligence and machine learning. But contrary to neural networks, genetic programming requires fewer prior restrictions to the structure of the model. Nevertheless, an often-cited disadvantage of both genetic programming and neural networks is the complexity of the developed prediction model ('black-box-character'). Genetic programming has been used in medical research for myoelectrical signal recognition, echocardiography and medical imaging but its value for medical prediction has not been documented yet.

Our aim was to compare genetic programming and multivariable logistic regression in the development of a diagnostic prediction model using empirical data from a study on diagnosis of pulmonary embolism (PE). We developed a prediction model using genetic programming and one using multivariable logistic regression, and compared both methods on their predictive accuracy in a validation set. The feasibility to apply both prediction models in clinical practice is discussed, as well as the differences between genetic programming and neural networks.

1_ Spiegelhalter DJ.
Probabilistic prediction in patient management and clinical trials.
Stat Med 1986;5:421-33.

2_ Hosmer D, Lemeshow S.
Applied logistic regression.
New York: John Wiley & Sons, Inc., 1989.

3_ Simon R, Altman DG.
Statistical aspects of prognostic factor studies in oncology.
Br J Cancer 1994;69:979-85.

4_ Harrell FE, Lee KL, Mark DB.
Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros.
Stat Med 1996;15:361-87.

5_ Laupacis A, Sekar N, Stiell IG.
Clinical prediction rules. A review and suggested modifications of methodological standards.
JAMA 1997;277:488-94.

6_ Harrell FE.
Regression modeling strategies.
New York: Springer-Verlag, 2001.

7_ Moons KG, Grobbee DE.
Diagnostic studies as multivariable, prediction research.
J Epidemiol Community Health 2002;56:337-8.

8_ Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB.
A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients.
J Investig Med 1995;43:468-76.

9_ Tsien CL, Fraser HS, Long WJ, Kennedy RL.
Using classification tree and logistic regression methods to diagnose myocardial infarction.
Medinfo 1998;9 Pt 1:493-7.

10_ Tu JV.
Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.
J Clin Epidemiol 1996;49:1225-31.

11_ Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R.
A comparison of statistical learning methods on the Gusto database.
Stat Med 1998;17:2501-8.

12_ Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV.
Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke.
J Clin Epidemiol 2001;54: 1159-65.

13_ Resnic FS, Ohno-Machado L, Selwyn A, Simon DI, Popma JJ.
Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention.
Am J Cardiol 2001;88:5-9.

14_ Holland JH.
Adaptation in Natural and Artificial Systems.
Ann Arbor: University of Michigan Press, 1975.

15_ Goldberg DE.
Genetic Algorithms in Search, Optimization and Machine Learning.
Addison Wesley Publishing Company, 1989.

Patients and Methods

Patients: description of the empirical data set

For the present analysis, data were used from a prospective diagnostic study among 398 patients in secondary care of 18 years or older who were suspected of PE. As data are used for illustration purposes only, we refer to literature for details on the design and main results of the study[18-20]. Briefly, all patients underwent a systematic patient history and physical examination, followed by blood gas analysis, chest radiography, leg ultrasound, ventilation-perfusion lung scanning (VQ-scanning) and pulmonary angiography. Chest X-ray was considered abnormal, i.e. indicative for presence of PE, if it showed an elevated hemidiaphragm, a small pleural effusion, atelectasis, consolidation or signs of heart failure. Leg ultrasound was considered abnormal if the femoral vein and/or popliteal vein were non-compressible. PE was considered present in case of a high probability VQ-scan or abnormal angiogram after a nonconclusive VQ-scan, and absent in case of a normal perfusion scan or normal angiogram after a nonconclusive VQ-scan. Of the 398 patients, 170 had PE (prevalence = 43%). All VQ-scans and angiograms were evaluated without knowledge of any other diagnostic information.

For our study, we a priori selected 10 candidate predictors, based on previous diagnostic studies[21-23]. These 10 predictors included 8 patient history and physical examination predictors, i.e. age, any co-existing malignancy, surgery within past three months, previous deep venous thrombosis, history of collapse, respiratory frequency, pleural rub, signs of deep venous thrombosis, and two from additional testing, i.e. abnormal leg ultrasound and abnormal chest X-ray. We note that these data are used for illustration purposes and not so much to report the optimal model for prediction of presence of PE to be used in future practice.

Methods

The data set was split, randomly, in two parts: a derivation set of approximately 67% (265 patients) and a validation set of approximately 33% (133 patients). The derivation set was used for model development (both by the logistic regression and genetic programming method) and the validation set to test the validity of the two models. The aim of both methods was to develop a prediction model to estimate the presence or absence of PE as good as possible with a minimum of diagnostic tests (predictors).

Multivariable logistic regression

In the derivation set, we first fitted the overall model including all 10 predictors. To enhance the flexibility of the logistic model and to obtain a more fair comparison with the (unrestricted) genetic programming, continuous variables (i.e. age and respiratory frequency) were included using cubic spline functions, both with 4 knots[4;6]. A reduced prediction model was obtained by selecting predictors with p-values <0.10 using the likelihood ratio test. To further enhance a fair comparison with genetic programming, we quantified whether interaction terms between the selected predictors increased the model's predictive accuracy. As our aim was not so much to develop an easy applicable model for future practice, we analysed many possible interaction terms. These, however, were not included all together but rather consecutively following the chronological order in which predictors are measured in practice and following

the order of contribution to the prediction[17]. Accordingly, we analyzed the interaction terms between chest X-ray with each of the selected history and physical examination predictors and of leg ultrasound with each of these predictors. In this, interaction terms of the selected history and physical predictors with the highest odds ratios were included first. The final model included all predictors and interaction terms with p-value <0.10.

Internal validation of the final model was performed using bootstrapping techniques[4;24]. Random samples were drawn with replacement from the derivation set with 100 replications, and the backward exclusion of the predictors including interaction terms was repeated within each bootstrap sample. Bootstrapping yielded an estimate of the overoptimism of the final model in predictive performance as expressed by the area under the receiver operating characteristic (ROC) curve[4;24;25]. Furthermore, a shrinkage factor was derived from the bootstrap samples to re-calibrate the model. To adjust the model for overoptimism, this factor was used as a shrinkage factor or multiplier of the regression coefficients of the predictors in the final model[4;24-26]. The re-calibrated model was applied to the validation set to estimate its discrimination and reliability in an independent sample. All analyses were performed using S-plus 2000 (Insightful Corp., Seattle WA, USA).

Genetic programming

Genetic programming is a search method inspired by the biological model of evolution[16;27]. It is an extension of the genetic algorithm first described by Holland[14] and Goldberg[15]. For the present analyses, we used the method of the OMEGA predictive modeling engine (KiQ Ltd., Cambridge, UK)[28] to search for a model that achieves optimal accuracy in predicting the presence or absence of PE.

In the genetic programming method by OMEGA, a prediction model is a mathematical formula, without inherent restrictions of complexity such as in logistic regression modeling that uses all predictors (or a subset of these) as inputs. The building blocks of the formula are mathematical operators, chosen from a library of 20 operators. Each operator has two inputs and one output (Figure 1, upper part). The output of the formula is a score, which is used to predict the presence of the outcome under study, a higher score indicating a higher probability. The fit of the formula is also expressed by the ROC area.

In the present study, first a set of 40 different prediction models was randomly created. This set consisted of 40 different mathematical formulas using different predictors. Then in an iterative process:

1. the fit of each model was determined by comparing the scores of the model with the observed PE frequencies in the derivation set;
2. various models were selected where models with a larger fit had a higher probability of being selected;
3. cross-over and mutation between the selected models occurred, creating new models;
4. these newly created models were moved to the next set of models and upon completing this set the next iteration started.

As the mathematical formulas or models all consist of binary operators, they can be represented as a binary tree (see Figure 1, upper part). To limit the amount of overoptimism, the trees were restricted to be no more than 4 levels deep, corresponding to a maximum of 8 predictors.

16_ Koza JR.
Genetic Programming III.
Cambridge, Massachusets: MIT Press, 1999.

17_ Knottnerus JA.
Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables.
Med Decis Making 1992;12: 93-108.

18_ van Beek EJ, Kuyer PM, Schenk BE, Brandjes DP, ten Cate JW, Büller HR.
A normal perfusion lung scan in patients with clinically suspected pulmonary embolism: frequency and clinical validity.
Chest 1995;108:170-3.

19_ van Beek EJ, Kuijer PM, Büller HR, Brandjes DP, Bossuyt PM, ten Cate JW.
The clinical course of patients with suspected pulmonary embolism.
Arch Intern Med 1997;157: 2593-8.

20_ Turkstra F, Kuijer PM, van Beek EJ, Brandjes DP, ten Cate JW, Buller HR.
Diagnostic utility of ultrasonography of leg veins in patients suspected of having pulmonary embolism.
Ann Intern Med 1997;126:775-81.

21_ Miniati M, Prediletto R, Formichi B, Marini C, Di Ricco G, Tonelli L et al.
Accuracy of clinical assessment in the diagnosis of pulmonary embolism.
Am J Respir Crit Care Med 1999;159:864-71.

22_ Stollberger C, Finsterer J, Lutz W, Stoberl C, Kroiss A, Valentin A et al.
Multivariate analyses-based prediction rule for pulmonary embolism.
Thrombosis research 2000;97:267-73.

23_ Wells PS, Ginsberg JS, Anderson DR, Kearon C, Gent M, Turpie AG et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Ann Intern Med 1998;129:997-1005.

24_ Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall, 1993.

25_ Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54: 774-81.

26_ Houwelingen van JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303-25.

27_ Banzhaf W, Nordin P, Keller RE, Francone FD. Genetic Programming, An Introduction. San Fransisco: Morgan Kaufmann Publishers Inc., 1998.

28. KiQ Ltd. http://www.kiq.com.

In step 3 new models are created by simulating the natural processes of sexual recombination (here called cross-over) between two chromosomes and mutation of DNA on a particular chromosome. In the context of genetic programming a prediction model or binary tree, as selected in step 2, can be compared with a chromosome. Here, cross-over and mutation operate on the branches (i.e. parts of the formula having one or more predictors as input, see Figure 1) and nodes of these trees. Given two selected models, cross-over is realised by swapping branches between the two trees. The swapped branches are randomly chosen. A mutation occurs by exchanging a node or a branch in a tree with a randomly created substitute, also here the node or branch that is mutated is randomly chosen. This random process, in addition to the probabilistic nature of the selection step (step 2), prevents the search method from converging at a local optimum.

The iterative process as mentioned above was terminated when no significant model improvement was observed. The model in the population with the largest ROC area was then selected as the final genetic programming model. A model provides a score for each patient in the data set. This score was transformed to a probability of PE presence by linking them to observed or actual proportions of PE.

Similar to the logistic regression method, the entire process of model development for genetic programming, including predictor selection, was bootstrapped to estimate the amount of overoptimism in predictive performance (ROC area). Because the formula developed by genetic programming has a different structure and is more complex, there are no regression coefficients estimated as compared to logistic regression models. Therefore, no shrinkage factor could be estimated and the final model could not be re-calibrated. The final genetic programming model was then also applied to the validation set to assess its discrimination and reliability in an independent sample.

Comparison of both methods

The two prediction models, obtained from multivariable logistic regression and genetic programming, were compared at their discrimination and reliability (calibration) in the validation set. Discrimination of both models was expressed by the area under the ROC curve. Reliability was evaluated by a graphical plot of the predicted probabilities versus observed or actual proportions of the outcome and tested using the Hosmer- Lemeshow test[2]. For logistic regression, per patient the predicted probability was calculated using the re-calibrated model, rank ordered and divided into deciles. For each decile the mean predicted probability and observed proportion of PE was calculated. For testing reliability of the genetic programming model, the output scores were rank ordered and also divided into deciles. For each decile the average observed PE proportion was calculated. As genetic programming yields a score and not directly a predicted probability, we used the observed proportion of PE in the derivation set as predicted probability for the calibration curve. Hence, the calibration curve of the genetic programming model compares the observed proportion of PE in the derivation set (x-axis) to those found in the validation set (y-axis). To enable comparison of the unadjusted calibration curves of both models, for logistic regression the predicted probability of PE was also plotted against the observed proportion before bootstrap shrinking.

As said, bootstrapping of the developed genetic programming model did not yield a shrinkage factor such that the model could not be adjusted for overfitting (re-calibration), in contrast to the logistic regression model. Hence, we also estimated the calibration curve of the original logistic model (before adjustment of overoptimism) and the genetic programming model for a fair comparison.

Results

Descriptives

There were no major differences in patient characteristics between the derivation and validation set (Table 1). PE was diagnosed in 42.6% of the patients in the derivation set, which was 42.9% in the validation set. Table 2 shows the univariable associations and distribution of the 10 predictors across patients with and without PE in the derivation set. 'History of collapse' and 'previous deep venous thrombosis' were the strongest predictors.

Model derivation

Logistic Regression

The overall logistic model yielded a ROC area of 0.77 (95% CI: 0.71-0.83) and the reduced model, including 8 predictors (Table 3), 0.76 (95% CI: 0.70-0.82). A restricted cubic spline transformation on age and respiratory frequency showed that the non-linear terms for both predictors were far from significant (p-value >0.40). Hence, age and respiratory frequency were analyzed as linear terms. Subsequently we consecutively added the interaction terms between leg ultrasound with each of the selected six history and physical predictors. In this, we first added the interaction term with collapse (as these showed the highest independent contribution to the prediction in predictive accuracy, Table 3), followed by signs of DVT, pleural rub, and so on. The same was done for chest X-ray. All interaction terms were far from significant (p-value> 0.40) except for age with chest X-ray, which was borderline significant. However, this interaction term did not increase the discriminative power of the reduced model at all. Therefore, the model presented in Table 3 was considered as the final logistic model.

Bootstrapping estimated the overoptimism at 0.06 in ROC area. Hence, the internally validated ROC area of the final model became 0.70. The bootstrap shrinkage factor for the regression coefficients was 0.76. Table 3 (column 2) shows the adjusted association (shrunk coefficients) of each predictor retained in the final model with the outcome.

Genetic Programming

The final model developed by genetic programming included 7 predictors (Figure 1). The ROC area of this model was 0.79 (95%CI: 0.73-0.85). Bootstrapping showed an estimated overoptimism of 0.07, decreasing the ROC area to 0.72.

Model validation

Applying the shrunk (adjusted) logistic regression model to the validation set yielded a ROC area of 0.68 (95% CI: 0.59-0.77), which was in good agreement to the ROC area estimated after bootstrapping. Application of the final genetic programming model to the validation set resulted in a ROC area of 0.73 (95% CI: 0.64-0.82), which was higher than the ROC area of the logistic model. Before shrinkage both models showed similar calibration curves and predicted rather accurately over the entire range of observed proportions of PE (Figure 2a and 2b). The Hosmer-Lemeshow test statistic was far from significant for both models (p-value >0.50), indicating good reliability. As expected, the reliability of the re-calibrated logistic model was better (Figure 2c).

Table 1. Comparison of predictors and the outcome between the derivation and validation set.

	Derivation set (n=265)	Validation set (n=133)
Age (years)	56.7 (17.8)*	53.8 (16.6)*
Any co-existing malignancy (%)	23.8	22.6
Surgery within past three months (%=)	21.1	21.1
Previous DVT (%)	6.4	9.8
History of collapse (%)	7.2	9.0
Respiratory frequency (breaths/min)	19.7 (6.7)*	18.2 (6.1)*
Pleural rub (%)	14.7	18.1
Signs of DVT (%)	8.7	10.5
Abnormal chest X-ray (%)	40.4	39.1
Abnormal leg ultrasound (%)	24.2	21.1
Pulmonary embolism present (%)	42.6	42.9

DVT = deep venous thrombosis; min = minute
* Mean (standard deviation)

Table 2. Univariable association of each predictor with the presence of pulmonary embolism in the derivation set (n=265).

Predictor	PE present (n=113)	PE absent (n=152)	Odds Ratio (95% CI)
Age (years)	60.6 (16.5)*	53.7 (18.3)*	1.02 (1.01-1.04)
Any co-existing malignancy (%)	31.0	18.4	2.0 (1.1-3.5)
Surgery within past three months (%)	29.2	15.1	2.3 (1.3-4.2)
Previous DVT (%)	10.6	3.3	3.5 (1.2-10.2)
History of collapse (%)	14.2	2.0	8.2 (2.3-28.9)
Respiratory frequency (breaths/min)	21.2 (7.5)*	18.5 (5.9)*	1.06 (1.02-1.11)
Pleural rub (%)	19.5	11.2	1.9 (0.97-3.8)
Signs of DVT (%)	13.3	5.3	2.8 (1.1-6.7)
Abnormal chest X-ray (%)	49.6	33.6	1.9 (1.2-3.2)
Abnormal leg ultrasound (%)	33.6	17.1	2.5 (1.4-4.4)

DVT = deep venous thrombosis; min = minute
* Mean (standard deviation)

Table 3. **Association of each predictor in the calibrated final logistic regression model with pulmonary embolism.**

Predictor	Odds Ratio	Regression Coefficient	P-value
Age (per year)	1.01	0.011	0.071
Surgery within past three months	1.57	0.45	0.090
History of collapse	5.01	1.61	0.002
Respiratory frequency (per breath/min)	1.05	0.044	0.013
Pleural rub	1.81	0.60	0.053
Signs of DVT	2.05	0.72	0.061
Abnormal leg ultrasound	1.98	0.68	0.006
Abnormal chest X-ray	1.46	0.38	0.096
Intercept		-2.46	

DVT = deep venous thrombosis
Probability of pulmonary embolism in an individual patient:
1/(1+exp-(-2.46 + 0.011*age + 0.45*surgery within past three months + 1.61*history of collapse + 0.044* respiratory frequency + 0.60*pleural rub + 0.72* signs of DVT + 0.68*abnormal leg ultrasound + 0.38*abnormal chest X-ray))

Model presentation

Logistic regression

The logistic model can be used in practice to estimate the probability of PE presence for individual patients in two different ways. First, one can use the formula as given in Table 3. With this formula one multiplies the patients' test results and corresponding coefficients, summing them and antilog the sum. This method, however, requires a calculator. An easier method is using a nomogram as presented in Figure 3. As an example of using this nomogram, a patient of 52 years of age (which corresponds to 4 points as determined from the 'Points' scale on top of the figure), with recent surgery (4 points), history of collapse (15 points), breathing frequency of 20 breaths per minute (4 points), no pleural rub (0 points), no signs of DVT (0 points), normal leg ultrasound (0 points) and an abnormal chest X-ray (4 points) receives a 'Total Points' score of 31. Using the lower two scales of the nomogram, this score corresponds to a probability of PE of approximately 0.8. The length of the line of each predictor in the nomogram also indicates the relative contribution of the predictor to the probability of PE.

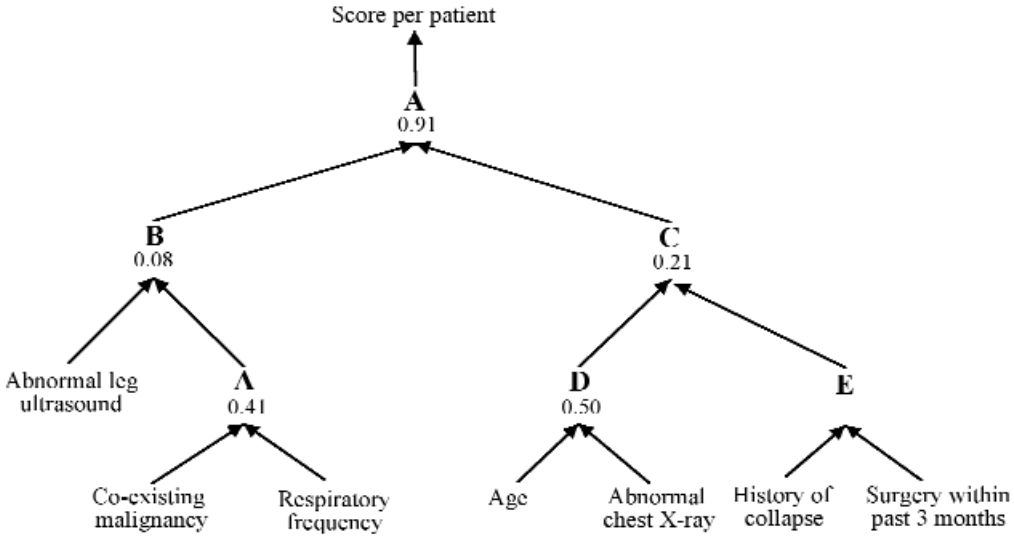
Genetic Programming

The model developed by genetic programming is shown in Figure 1. The same patient as described above is taken as an example to calculate the probability of PE presence using the genetic programming model. The values of the predictors except for leg ultrasound (age= 52 years, surgery in past 3 months=1, co-existing malignancy=1, history of collapse=1, respiratory rate=20 breaths per minute, pleural rub=0, abnormal chest X-ray=1) are the inputs (x and y) for the boxes A, D and E. The outputs of these boxes (using the formulas per box shown in the figure) plus abnormal leg ultrasound=0, are then used as inputs of box B and C which in turn yield the input for box A. Box A results in the final score for the patient which was 9.05. The score table in Figure 1 shows that this score corresponds to a probability of PE of 0.79, which was similar to the probability obtained by the logistic regression model.

Discussion

To our knowledge, this is the first study to address the value of genetic programming for medical prediction purposes as compared to the well-known and widely applied logistic regression technique. Given that the amount of overoptimism in discriminative value was similar for both models as estimated from the bootstrap, the discriminative value of the genetic programming model in the validation set was significantly larger than that of the logistic regression model. Before any form of re-calibration or adjustment for overoptimism of the logistic model, the reliability in the validation set was similar for both models. However, the logistic model showed improved reliability after it was beforehand re-calibrated through shrinkage. These results indicate that genetic programming offers a promising technique for prognostic and diagnostic prediction research, in particular when the aim is to achieve optimal discrimination. To appreciate the results a few issues need to be addressed.

Because of the more complex structure of a genetic programming model, it does not provide regression coefficients or odds ratios that indicate the relative predictive contribution of each predictor. Therefore, logistic regression techniques remain first choice when the primary goal of an epidemiological study is to examine the (relative) strength of the association between risk factors and the outcome,



Score category	Probability of pulmonary embolism
0.00-5.80	0.15
5.81-6.90	0.21
6.91-7.00	0.43
7.01-8.00	0.59
8.01-10.0	0.79

Figure 1. The final model created by genetic programming, presented as a binary tree and output scores from the tree related to the observed proportion of the outcome. The nodes A-E represent the following binary operators, in which the parameters x (left arrow) and y (right arrow) are the inputs of each box

$$\mathrm{A} = 1 - p\sqrt{(1-x)} - (1-p)\sqrt{(1-y)}$$

$$\mathrm{B} = p\,\mathrm{f}(x) + (1-p)\,\mathrm{f}(y) \quad \text{where } \mathrm{f}(x) = 2x - k(2x-1)^3 \text{ and } k = 0.593$$

$$\mathrm{C} = px + (1-p)y$$

$$\mathrm{D} = px^2 + (1-p)y^2$$

$$\mathrm{E} = \tfrac{1}{2} + \tfrac{1}{2}\sin(x^2 + \tfrac{1}{2}\pi y^2 - 1).$$

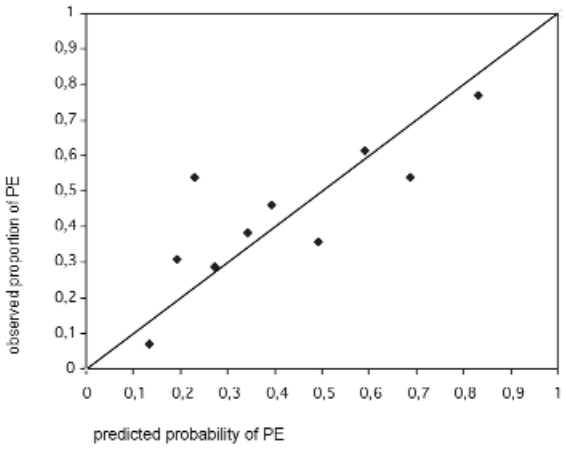


Figure 2a. Calibration curve of the original (without application of bootstrap shrinkage) logistic regression model in the validation set.

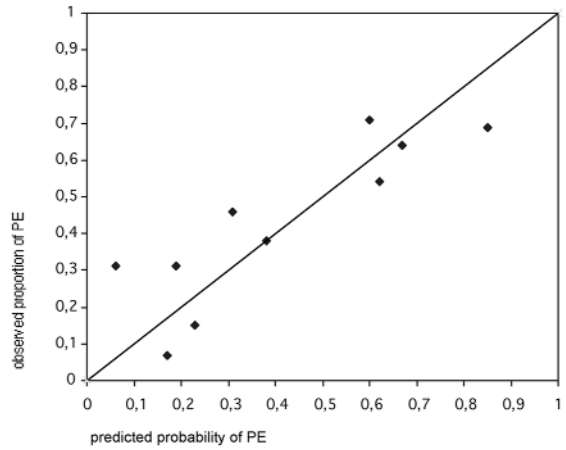


Figure 2b. Calibration curve of the genetic programming model in the validation set.

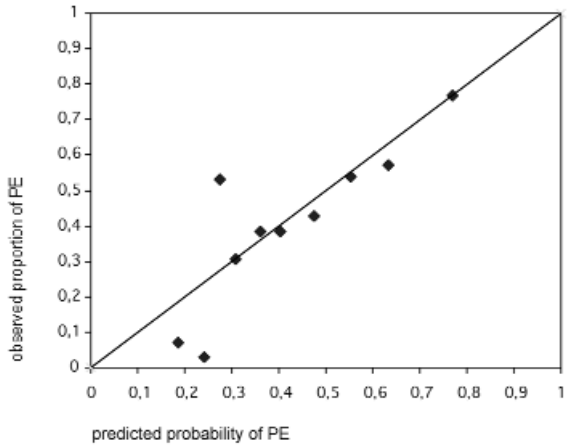


Figure 2c. Calibration curve of the re-calibrated (i.e. after application of bootstrap shrinkage) logistic regression model in the validation set.

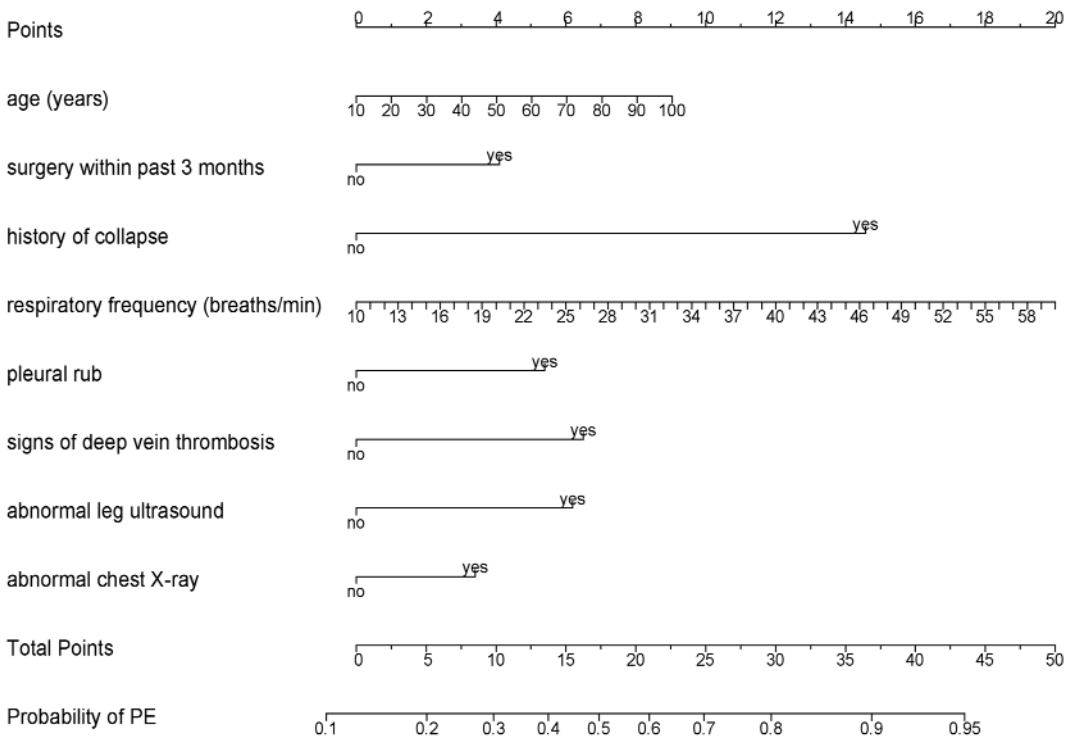


Figure 3. Nomogram relating the predictors of the final model (Table 3) to the probability of pulmonary embolism. See text for instructions on how to use a nomogram.

such as in etiologic research. However, when the aim is to obtain an optimal prediction for individual patients in clinical practice, as is common for diagnostic and prognostic studies, researchers might search for more complex models as developed by genetic programming.

Due to the more complex nature of genetic programming, the final model is often less intuitive and interpretable ('black box' character). Some degree of insight can be given into the functioning of a genetic programming model by studying the sensitivity of the model to the used predictors, thereby identifying predictors that have a large influence on the output scores or discriminative value of the model. Also, insight can be gained by describing profiles of predictor values in the different score intervals and thus to specific risk categories. Nevertheless, the complexity of prediction rules developed by genetic programming may influence their use in medical practice. Application of a genetic programming model for prediction purposes in practice requires a personal computer equiped with the necessary software. Prediction rules developed by logistic regression can be easier disseminated since they directly yield a predicted probability by either using the regression coefficients and a pocket calculator or using a nomogram as presented in Figure 3. However, with proceeding computerization in clinical practice and the rise of electronic patient records this difference in applicability between both methods may become irrelevant.

The main advantage of the complexity and fewer restrictions of genetic programming, is the possibility to create more flexible prediction models with better discrimination as was also exemplified in our analyses. This may be of even more importance in larger data sets in which complex interactions between predictors and outcomes may be present, although such interactions can also be modeled in a logistic model.

As a prediction model developed by genetic programming does not contain regression coefficients, a method of re-calibration through shrinkage of these coefficients using data from the derivation set only (so-called internal validation) is not possible, in contrast to logistic regression. Furthermore, the patient score estimated by a genetic programming model is not linearly related to a probability of the outcome. To relate a score to probabilities, the observed frequency of the outcome must be estimated for different score intervals. These frequencies (probabilities) can be obtained from the data set from which the model is derived (derivation set). Subsequently, to determine its reliability, the genetic programming model should be applied to a validation set to compare the observed outcome frequencies per score interval from the derivation set to those observed in the validation set. To prevent overoptimism of the genetic programming model in practice, the probabilities from the validation set should then be used and presented as the probabilities that might be expected in future subjects. This method of re-calibration for genetic programming models (as also done in our study) is appealing, but requires the use of a data set that is large enough to perform a split sample method.

An advantage of logistic regression is that a developed prediction model can easily be re-calibrated using data from the derivation set only, e.g. using shrinkage with the bootstrapping method (as done in our study), before it is to be applied to future patients. This is of particular interest when the available data to develop a prediction model is scarce. In our example study, a priori re-calibration of the logistic model using internal validation techniques, indeed improved the reliability of the model in 'new' patients (Figure 2c). A similar technique of re-calibration with the bootstrapping method using data of the derivation set only can also be used

for genetic programming, although without shrinking of coefficients. Instead, the probabilities per score interval are re-calibrated by taking the average of the observed frequencies per interval obtained for each bootstrap model. This is a relatively time-consuming method since each bootstrap set requires the development of a new model. However, this method was performed and checked in our example study, and produced similar results to the validation method of re-calibration. We only presented the latter results for reasons of clarity and because frequencies (probabilities) from another patient (validation) set better reflect future practice.

Finally, genetic programming requires a number of parameters to be chosen by the researcher, such as the number of prediction models that are evolved in parallel (we used 40 models), the selection method and the probabilities of crossover and mutation. Although the setting of these parameters requires some experience and certainly influences the speed of the search method, the final result was not very sensitive to these parameters; using other parameter settings did not result in other discrimination and reliability of the final model. The maximum depth for each tree is another parameter to be chosen by the researcher. It is chosen to limit the number of degrees of freedom and is similar to the number of predictors that would be included in the final logistic regression model. Allowing for more degrees of freedom commonly results in a more overoptimistic prediction model.

Since neural networks are a well known method and have also been used in the medical field to produce non-linear prediction models, it is interesting to briefly discuss the differences between neural networks and genetic programming. Although the representation of the model created by genetic programming as shown in Figure 1 may at first sight show similarities to a neural network model, it must be noted that only the representation in the form of a tree of the finally obtained (mathematical) formula is similar. In a genetic programming model, there are no hidden nodes. Instead the boxes in the model (see Figure 1) represent parts of the mathematical formula that are called operators. Since in this case the chosen operators are all binary the representation is a binary tree. Indeed, a neural network as well is a presentation of a mathematical formula, but of a very specific and pre-determined form. Given the chosen set of operators, the possible predictors to choose from, and restrictions on the number of degrees of freedom, the method of genetic programming is free to construct a more optimal mathematical prediction formula. The training period of a neural network (or the log-likelihood fit of a logistic regression), is in the case of genetic programming replaced by an iterative search process in which large numbers of possible models (mathematical formulas of different form, using different predictors) are evaluated in parallel and which finally evolves into the most optimal solution.

In conclusion, using empirical data we demonstrated that a prediction model developed by the novel technique of genetic programming may have an increased discriminative power with comparable reliability, compared to a model developed by logistic regression. Although this is the first empirical study quantifying the value of genetic programming for medical prediction and more empirical studies are needed, it seems a promising technique to develop prediction rules for diagnostic and prognostic purposes.

chapter
7

Revisiting polytomous regression for diagnostic studies

Introduction

Diagnostic practice starts with a patient presenting with particular signs and symptoms. The physician soon defines the differential diagnoses and, although implicitly, estimates the probability of presence of each of these possible diseases given the patient's clinical and non-clinical profile[1-3]. Usually, one of these differential diagnoses is defined as the working diagnosis or target disease, to which the diagnostic work-up is primarily directed. Consequently, diagnostic test evaluation studies commonly focus on the ability of tests to include or exclude the presence or absence of this target disease. The alternative diagnoses (which may all direct different treatment decisions) are thus included in the outcome category 'target disease absent'. Similarly, scientific studies that aim to develop so-called diagnostic prediction rules use dichotomous logistic regression analysis to predict the presence or absence of the target disease. Well known examples are the Ottawa ankle rule to diagnose ankle fracture[4] and the Wells rule to diagnose deep venous thrombosis[5]. However, diagnostic prediction rules developed with dichotomous logistic regression, simplify clinical practice. A rule to estimate the probabilities of presence of each of the potential diseases given the patient's characteristics or test results, may serve diagnostic practice better.

Already in the early eighties Wijesinha et al and Gray et al discussed the use of polytomous logistic regression to accommodate simultaneous prediction of more than two unordered outcome categories[6;7]. However, this method has received little attention for diagnostic purposes. We believe it is timely to revisit polytomous regression to address diagnostic questions.

We provide an introduction into the principles of polytomous logistic regression and will illustrate its utility using empirical data from a study on diagnosing residual retroperitoneal mass histology in patients with nonseminomatous testicular germ cell tumour (NSTGCT)[8]. The differential diagnoses in these patients include benign tissue, mature teratoma, and viable cancer. The accuracy of a polytomous model to estimate the probability of each of the three diagnoses will be compared with that of two consecutive dichotomous logistic regression models, in which one model aims to discriminate benign tissue from the other two histologies and a second model aims to discriminate mature teratoma from viable cancer. Finally, the advantages and disadvantages of polytomous logistic regression are discussed.

Patients and Methods

Patients

For the present analyses we used data of previous studies on residual retroperitoneal mass histology in patients treated with chemotherapy for metastatic NSTGCT[8-11]. These studies were performed to develop and validate a dichotomous diagnostic model to discriminate benign tissue from other histologies. The combined studies included 1094 patients treated in various hospitals across different countries. Patients with elevated levels of the serum tumour markers alpha-fetoprotein (AFP) or human chorionic gonadotropin (HCG) at the time of surgery, extragonadal primaries, histological pure seminoma without elevated prechemotherapy serum tumour markers, or resection after relapse were excluded.

The differential diagnoses in patients with suspected residual masses after chemotherapy for NSTGCT include besides benign tissue and viable cancer, the diagnosis of mature teratoma. Surgical resection is a generally accepted treatment

1_ Feinstein AR.
Clinical Epidemiology: the architecture of clinical research.
Philadelphia: WB Saunders Company, 1985.

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

3_ Moons KG, Grobbee DE.
Diagnostic studies as multivariable, prediction research.
J Epidemiol Community Health 2002;56:337-8.

4_ Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR.
A study to develop clinical decision rules for the use of radiography in acute ankle injuries.
Ann Emerg Med 1992;21: 384-90.

5_ Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C et al.
A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process.
J Intern Med 1998;243:15-23.

6_ Wijesinha A, Begg CB, Funkenstein HH, McNeil BJ.
Methodology for the differential diagnosis of a complex data set. A case study using data from routine CT scan examinations.
Med Decis Making 1983;3: 133-54.

7_ Begg CB, Gray D.
Calculation of polychotomous logistic regression parameters using individualized regressions.
Biometrika 1984;71:11-8.

to remove the residual masses. However, if the likelihood is high that the mass contains only benign tissue, surgery may be withheld[8;12;13]. Therefore, the distinction between benign tissue and viable cancer is clinically most important[14-18]. It may also be relevant to differentiate mature teratoma from viable cancer and benign tissue pre-operatively. Patients with a high likelihood of mature teratoma may not require immediate surgery as patients with a high likelihood of viable cancer do, and may require closer follow-up than patients with a high likelihood of benign tissue. Hence, it is clinically relevant if one could simultaneously discriminate pre-operatively between benign tissue, mature teratoma, and viable cancer.

Predictors

In the present analyses, we included three dichotomous and three continuous predictors that were found the most important predictors in previous research[8;13;19]. Dichotomous predictors were the absence of teratoma elements in the primary tumour, and normal prechemotherapy levels of the serum tumour markers AFP and HCG. Continuous predictors included the standardised value of prechemotherapy level of the serum tumour marker lactate dehydrogenase (standardised value of LDH = LDH level /upper limit of the normal value), the maximum diameter (in mm) of the residual mass measured on computed tomography (CT) after chemotherapy (postchemotherapy mass size), and the reduction in mass size (per 10%) after chemotherapy.

Outcome

To determine the final diagnosis all patients underwent resection of the residual retroperitoneal mass of which histopathology was determined (reference standard). Viable cancer refers to masses that contained viable cancer cells and possibly mature teratoma or benign tissue. Mature teratoma refers to masses that contained mature teratoma and possibly benign tissue (necrosis/fibrosis), but no viable cancer cells. Benign tissue refers to masses without viable cancer and mature teratoma elements. The outcome thus included three categories, i.e. viable cancer, mature teratoma, and benign tissue.

Dichotomous versus polytomous logistic regression

Dichotomous logistic regression to estimate the probability that an outcome is present (versus absent) models the log odds (logit) of the outcome probability as a function of one or more predictors:

$$Log\ ((Probability\ (outcome\ event)\ /\ Probability\ (non\text{-}event)) = \beta_0 + \beta_1 X_1 + \ldots\ \beta_6 X_6 = linear\ predictor\ (lp) \quad (1)$$

in which β_0 to β_6 are regression coefficients of X_1 to X_6 which are the six predictors defined above. The sum of the regression coefficients multiplied by the predictor values is called the linear predictor. A regression coefficient can be interpreted as the log odds of the outcome event relative to a non-event per unit change in a specific predictor value. The odds ratio can be computed as the antilog of the regression coefficient. Formula 1 can be rewritten to estimate the probability of occurrence of the outcome event for each patient. In the example of predicting the presence of benign tissue versus other histology, i.e. the presence of mature teratoma or viable cancer, the formula is:

8_ Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Toner GC, Schraffordt Koops H et al. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol 1995;13:1177-87.

9_ Steyerberg EW, Gerl A, Fossa SD, Sleijfer DT, de Wit R, Kirkels WJ et al. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. J Clin Oncol 1998;16:269-74.

10_ Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol 2001;165:84-8.

11_ Vergouwe Y, Steyerberg EW, de Wit R, Roberts JT, Keizer HJ, Collette L et al. External validity of a prediction rule for residual mass histology in testicular cancer: an evaluation for good prognosis patients. Br J Cancer 2003;88:843-7.

12_ Steyerberg EW, Keizer HJ, Habbema JD. Prediction models for the histology of residual masses after chemotherapy for metastatic testicular cancer. ReHiT Study Group. Int J Cancer 1999;83:856-9.

13_ Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JD. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Stat Med 2001;20:3847-59.

*Probability (benign tissue) = exp(β_0+ β_1*X_1+... β_6*X_6) / (1 + exp(β_0+ β_1*X_1+... β_6*X_6)) = 1 / [1 + exp(lp)]* **(2)**

Accordingly, the probability of other histology can thus be estimated with:

Probability (mature teratoma or viable cancer) = 1- Probability (benign tissue) **(3)**

Polytomous logistic regression analysis is advocated when there are more than two unordered outcome categories. In the diagnostic context, when the differential diagnoses include more than two alternative diseases, one of the diagnostic outcome categories is chosen as reference category. For the other outcome categories the polytomous regression method simultaneously fits sub models that compare the outcome categories with the chosen reference[6;20]. Thus, for each outcome category, different regression coefficients are fitted. These sub models together contain the polytomous model and can be used to estimate the probability of presence of each diagnostic outcome. In our example study, the reference diagnosis was viable cancer. Hence, we fitted a polytomous regression model, consisting of two sub models, one for benign tissue compared to viable cancer, and one for mature teratoma compared to viable cancer. These models take a similar form as formula 1:

*Log (Probability 'benign tissue'/ Probability 'viable cancer') = βb_0 + βb_1*X_1+... βb_6*X_6 = lpb* **(4)**

*Log (Probability 'mature teratoma' / Probability 'viable cancer') = βt_0 + βt_1*X_1 +... βt_6*X_6 = lpt* **(5)**

The subscripts of the regression coefficients correspond with the outcome category (i.e. 'b' for benign tissue and 't' for mature teratoma). The interpretation of the regression coefficients is similar as for dichotomous logistic regression, i.e. the log odds of the outcome (benign tissue or mature teratoma) relative to viable cancer per unit change in the predictor values. The probability of benign tissue can be calculated by:

Probability (benign tissue) = exp(lpb) / [1 + exp(lpb) + exp(lpt)] **(6)**

The probability of mature teratoma can be calculated by:

Probability (mature teratoma) = exp(lpt) / [1 + exp(lpb) + exp(lpt)] **(7)**

As probabilities in logistic regression analysis always count up to 1, the probability of viable cancer can then be calculated by:

Probability (viable cancer) = 1 – Probability (benign tissue) - Probability (mature teratoma) **(8)**

14_ Fossa SD, Qvist H, Stenwig AE, Lien HH, Ous S, Giercksky KE.
Is postchemotherapy retroperitoneal surgery necessary in patients with nonseminomatous testicular cancer and minimal residual tumor masses?
J Clin Oncol 1992;10:569-73.

15_ Gelderman WA, Schraffordt KH, Sleijfer DT, Oosterhuis JW, Van der Heide JN, Mulder NH et al.
Results of adjuvant surgery in patients with stage III and IV nonseminomatous testicular tumors after cisplatin-vinblastine-bleomycin chemotherapy.
J Surg Oncol 1988;38:227-32.

16_ Toner GC, Panicek DM, Heelan RT, Geller NL, Lin SY, Bajorin D et al.
Adjunctive surgery after chemotherapy for nonseminomatous germ cell tumors: recommendations for patient selection.
J Clin Oncol 1990;8:1683-94.

17_ Donohue JP, Rowland RG, Kopecky K, Steidle CP, Geier G, Ney KG et al.
Correlation of computerized tomographic changes and histological findings in 80 patients having radical retroperitoneal lymph node dissection after chemotherapy for testis cancer.
J Urol 1987;137:1176-9.

18_ Mulders PF, Oosterhof GO, Boetes C, de Mulder PH, Theeuwes AG, Debruyne FM.
The importance of prognostic factors in the individual treatment of patients with disseminated germ cell tumours.
Br J Urol 1990;66:425-9.

19_ Steyerberg EW, Keizer HJ, Stoter G, Habbema JD. Predictors of residual mass histology following chemotherapy for metastatic non-seminomatous testicular cancer: a quantitative overview of 996 resections. Eur J Cancer 1994;30A:1231-9.

20_ Agresti A. An introduction to categorical data analysis. New York: John Wiley & Sons, 1996.

21_ Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 1988;80:1198-202.

22_ Harrell FE. Regression modeling strategies. New York: Springer-Verlag, 2001.

23_ Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

24_ Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247:2543-6.

25_ Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Methods Inf Med 1978;17: 238-46.

26_ Ash A, Shwartz M. R^2: a useful measure of model performance when predicting a dichotomous outcome. Stat Med 1999;18:375-84.

Data analysis

In previous analyses, the standardised value of LDH was transformed to the natural logarithm scale, postchemotherapy mass size was transformed with the square root and change in mass size was fitted linearly (per 10%) for the outcome benign tissue versus other histology[8;13]. For the present analysis, we re-studied the transformations of these continuous predictors, given the three outcome categories. We fitted restricted cubic spline functions for each continuous predictor with the outcome categories benign tissue versus viable cancer and mature teratoma versus viable cancer. The splines were then approximated with simple transformation[21;22].

We fitted a multivariable polytomous logistic regression model with the six predictors to enable estimation of the probability of benign tissue, mature teratoma and viable cancer. Variable selection was not applied.

For comparison reasons, we also fitted two consecutive multivariable dichotomous logistic models using the same six predictors. The first model aimed to discriminate benign tissue from the other two outcome categories (mature teratoma and viable cancer). The second, consecutive, model aimed to discriminate mature teratoma from viable cancer in patients who did not have benign tissue. With these two models for each patient the probability of presence of benign tissue (versus no benign tissue) was calculated by:

$$\textit{Probability (benign tissue)} = \exp(lpb) / 1 + \exp(lpb)) \quad \textbf{(9)}$$
$$\textit{in which } lpb = ßb_0 + ßb_1{*}X_1 + \ldots ßb_6{*}X_6 \textit{ as defined by formula 4}$$

The probability of mature teratoma could be calculated by:

$$\textit{Probability (mature teratoma)} = (1 - \textit{Probability (benign tissue)}) * (\exp(lpt) / [1 + \exp(lpt)]) \quad \textbf{(10)}$$
$$\textit{in which } lpt = ßt_0 + ßt_1{*}X_1 + \ldots ßt_6{*}X_6 \textit{ as defined by formula 5}$$

The probability of viable cancer could then be calculated according to formula 8.

Aspects of model performance

We studied several aspects of performance of the polytomous and two consecutive dichotomous models. These aspects included calibration, discrimination, overall performance, and diagnostic classification accuracy.

Calibration refers to the amount of agreement between predicted and observed outcomes. For instance, if patients with certain characteristics are predicted to have a 70% probability of benign tissue, then 70% of such patients should indeed have benign tissue at resection. A graphical impression of calibration was obtained for the polytomous model and the two consecutive dichotomous models, by plotting the observed frequencies with the predicted probabilities for each diagnostic outcome category versus the two other categories.

Discrimination refers to the ability to discriminate between the different diagnostic outcomes (in this study the presence of benign tissue, mature teratoma and viable cancer). Commonly, for dichotomous and ordinal logistic regression models the c-statistic or area under the receiver operating characteristics curve (ROC area) is used as single measure of discrimination[22-24]. However, if the outcome categories are unordered, as is the case with polytomous regression, discrimination cannot be estimated by a single ROC area. Therefore, we calculated three ROC areas, each time relating one outcome category versus the other two outcome categories. We

also calculated the ROC area for the two consecutive dichotomous models. The overall model performance was measured with the Brier score and R^2. The Brier score estimates the mean difference between the observed outcomes and the predicted probabilities[25]. This score ranges from 0 to 2 if more than two outcome categories are predicted. The lower the score, the better the performance. R^2 is interpretable as the proportion of the variation in the outcomes, which can be explained by the predictors in the model and runs from 0 to 1. The higher the R^2 value, the better the model performance[26].

We calculated the proportion of correctly classified patients using clinical relevant thresholds for assigning patients to one of the three outcome categories. The proportion of correctly classified patients was obtained from a three by three diagnostic outcome table. The rows of this table represent the diagnostic outcome categories as predicted by the polytomous or consecutive dichotomous models. The columns represent the diagnostic outcome categories as observed by histology (reference standard). Patients are correctly classified if the predicted diagnostic outcome category corresponds to the observed diagnostic outcome. Patients were assigned to a diagnostic category (the rows) using clinically relevant thresholds in the models' predicted probabilities. Presence of viable cancer was assumed to be eight times as worse as benign tissue as published previously[27], and the presence of mature teratoma was assumed to be three times as worse as benign tissue (benign tissue: mature teratoma: viable cancer = 1:3:8). Hence, patients were assigned to the viable cancer category, unless the predicted probability for benign tissue exceeded 0.89 (8/9) or the predicted probability for the mature teratoma outcome exceeded 0.75 (3/4). In those cases, patients were assigned to the benign tissue category or the mature teratoma category respectively. For example, a patient with predicted probabilities of 0.18, 0.70, and 0.12 for benign tissue, mature teratoma and viable cancer respectively, was assigned to the viable cancer category.

Analyses were performed using SPSS version 12.0 and S-plus version 6.2 software.

Model presentation

Finally, the polytomous model was presented as a score chart which facilitates estimation of individual patient probabilities for the three outcome categories in clinical practice. To derive the scores, the model's regression coefficients were multiplied by 5 and subsequently rounded. A constant was added to avoid negative scores where possible.

27_Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Bajorin DF, Donohue JP et al. Resection of residual retroperitoneal masses in testicular cancer: evaluation and improvement of selection criteria. The ReHiT study group. Re-analysis of histology in testicular cancer. Br J Cancer 1996;74:1492-8.

28_ Flack VF, Chang PC. Frequeny of selecting noise variables in subset regression: a simulation study. The American Statistician 1987;41:84-6.

29_ Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. Cancer Treat Rep 1985;69: 1071-7.

30_ Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.

31_ Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52: 935-42.

32_ Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21: 45-56.

33_ Marshall RJ, Chisholm EM. Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer. Stat Med 1985;4:337-44.

Table 1. Distribution of predictors across outcome categories and in the total study population (n = 1094).

Characteristic	Benign tissue N (%)	Mature teratoma N (%)	Viable cancer N (%)	Total data set N (%)
Predictors				
Primary tumour teratoma negative	279 (55)	170 (34)	54 (11)	503 (46)
Normal AFP level	200 (59)	112 (33)	27 (8)	339 (31)
Normal HCG level	184 (49)	154 (41)	40 (10)	378 (35)
Standardised value of LDH*	1.5 (0.39-69.7)	1.2 (0.12-20.9)	1.8 (0.34-63.6)	1.4 (0.1-70)
Postchemotherapy mass size (mm)*	18.0 (2.0-300)	30.0 (2.0-300)	40.5 (2.0-300)	28.2 (2.0-300)
Reduction in mass size after chemotherapy (%)*	60 (-150-100)	20 (-150-100)	43 (-250-100)	43 (-250 - 100)
Histology at resection	425 (39)	535 (49)	134 (12)	1094 (100)

* Median (range)
AFP alfa-fetoprotein
HCG human chorionic gonadotropin
LDH lactate dehydrogenase

Table 2. Results of the multivariable polytomous and consecutive dichotomous logistic regression analysis. Values represent odds ratios with corresponding 95% confidence intervals.

Predictor	Polytomous regression		Dichotomous regression	
	benign tissue vs. viable cancer	mature teratoma vs. viable cancer	benign tissue vs. other histologies	mature teratoma vs. viable cancer
Primary tumour teratoma negative	2.2 (1.4-3.3)	0.66 (0.44-0.99)	3.0 (2.2-4.0)	0.61 (0.40-0.92)
Normal AFP serum level	2.8 (1.7-4.6)	0.94 (0.57-1.5)	2.9 (2.1-4.0)	0.90 (0.54-1.5)
Normal HCG serum level	1.4 (0.89-2.3)	0.72 (0.46-1.14)	1.9 (1.3-2.6)	0.70 (0.44-1.1)
Transformed standardised value of LDH*	1.2 (0.84-1.6)	0.58 (0.42-0.78)	1.7 (1.4-2.2)	0.60 (0.44-0.81)
Transformed postchemotherapy mass size† (mm)	0.79 (0.71-0.88)	0.91 (0.84-0.99)	0.85 (0.77-0.92)	0.89 (0.82-0.98)
Reduction in mass size after chemotherapy (per 10%)	1.14 (1.06-1.22)	0.97 (0.92-1.02)	1.2 (1.1-1.2)	0.96 (0.92-1.0)

*Transformation: ln (LDHst)
†Transformation: square root (postchemotherapy mass size)

Results

In 425 (39%) out of 1094 patients the final diagnosis was benign tissue, 535 (49%) had mature teratoma, and 134 (12%) had viable cancer. Table 1 shows the distributions of the six predictors across the three diagnostic outcome categories and in the total study population. Overall, 46% of the patients had teratoma negative primary tumour histology. Tumour markers AFP and HCG were normal in approximately one third of all patients (31% and 35%, respectively). The natural logarithm of the standardised value of prechemotherapy LDH and the square root of residual mass size (in mm) were again accurate transformations to describe the associations with the diagnostic outcome categories. Reduction in mass size (per 10%) could be described with a linear term.

Polytomous model

Except postchemotherapy mass size, all odds ratios of predictors for benign tissue versus viable cancer were larger than 1.0 (Table 2). This indicates that primary tumour teratoma negative, normal levels of AFP and HCG, higher levels of LDH, and a larger reduction in mass size increase the odds of benign tissue as opposed to viable cancer. A larger postchemotherapy mass size decreases the odds of benign tissue. For example, the odds ratio of 2.2 for the predictor 'primary tumour teratoma negative' indicates that a patient with a teratoma negative primary tumour has a 2.2 times higher probability for benign tissue as opposed to viable cancer than a patient with a teratoma positive primary tumour. All odds ratios of predictors for mature teratoma versus viable cancer were smaller than 1.0, indicating that they all decrease the odds of mature teratoma as opposed to viable cancer.

Table 2 also shows the associations of the predictors with the outcomes for the two consecutive dichotomous models. The odds ratios of the predictors that discriminated benign tissue from mature teratoma and viable cancer combined in the first model were slightly different from the predictors in the polytomous model that discriminated benign tissue from viable cancer. The odds ratios of the predictors that discriminated teratoma from viable cancer in the second dichotomous model were very similar to those in the polytomous model.

Aspects of model performance

Calibration of the polytomous model is shown in Figure 1. Overall, the correspondence between predicted probabilities and observed frequencies was good for all three outcome categories. Calibration plots for the three outcome categories estimated with the consecutive dichotomous models were similar to those of the polytomous model (Figure 2).

The values of the ROC areas for the polytomous model and the two consecutive dichotomous models are presented in Table 3. The ROC areas for benign tissue (0.83) and mature teratoma (0.78) were the same for both methods. The ROC area of viable cancer in the consecutive dichotomous models was slightly, but not significantly, lower than the ROC area of the polytomous model (0.64 versus 0.66).

For the polytomous model, the Brier score was 0.458 and the R^2 was 0.392 (Table 3). The overall performance of the consecutive dichotomous models was slightly lower as shown by a higher Brier score (0.461) and a lower R^2 (0.345).

For the polytomous model, the proportion of detected (correctly classified) viable cancer cases was 68% (91/134), 36% (193/535) for mature teratoma cases and 36% (155/425) for patients with benign tissue (Table 4a). For the consecutive dichotomous models, the proportions of detected viable cancer and benign tissue

34_ Harrell FE, Lee KL, Mark DB.
Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros.
Stat Med 1996;15:361-87.

35_ Jaeschke R, Guyatt GH, Sackett DL.
Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid?
JAMA 1994;271:389-91.

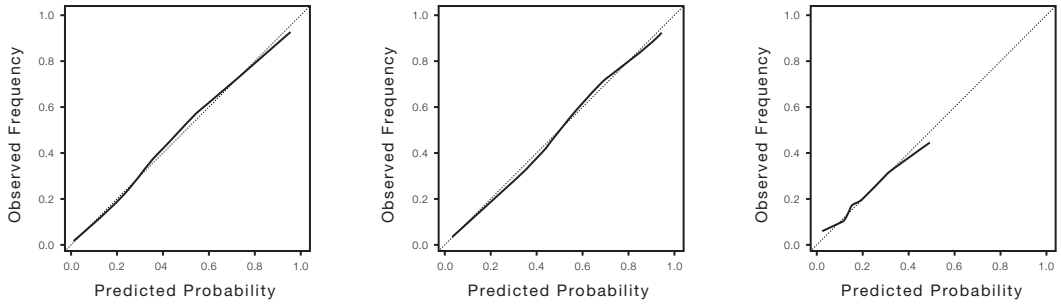
36_ Justice AC, Covinsky KE, Berlin JA.
Assessing the generalizability of prognostic information.
Ann Intern Med 1999;130: 515-24.

37_ Wasson JH, Sox HC, Neff RK, Goldman L.
Clinical prediction rules. Applications and methodological standards.
N Engl J Med 1985;313:793-9.

38_ Efron B, Tibshirani R.
Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy.
Statistical Science 1986;1: 54-77.

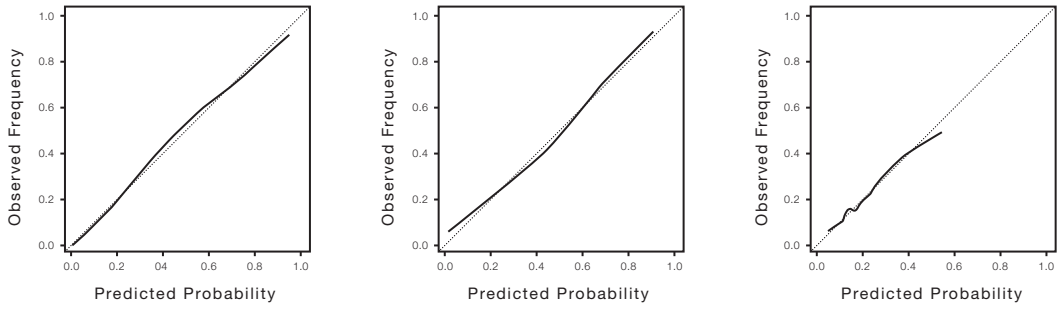
39_ Efron B, Tibshirani R.
An introduction to the bootstrap. Monographs on statistics and applied probability.
New York: Chapman & Hall, 1993.

40_ Houwelingen van JC.
Shrinkage and penalized likelihood methods to improve diagnostic accuracy.
Stat Neerl 2001;55:17-34.



a. Benign tissue b. Mature teratoma c. Viable cancer

Figure 1. Calibration plots of the polytomous model for the three outcome categories. The solid line shows the smoothed association between the predicted probabilities and observed frequencies. Ideally, this line equals the dotted line.



a. Benign tissue b. Mature teratoma c. Viable cancer

Figure 2. Calibration plots of the two consecutive dichotomous models for the three outcome categories. The solid line shows the smoothed association between the predicted probabilities and observed frequencies. Ideally, this line equals the dotted line.

Table 3. Performance parameters of the polytomous and consecutive dichotomous models. ROC areas are presented with corresponding 95% confidence intervals.

Parameter	Model	
	Polytomous	Consecutive dichotomous
ROC area[1]	0.83 (0.80-0.85)	0.83 (0.80-0.85)
ROC area[2]	0.78 (0.75-0.81)	0.78 (0.75-0.81)
ROC area[3]	0.66 (0.61-0.71)	0.64 (0.59-0.69)
Brier	0.458	0.461
R^2	0.392	0.345

[1] Benign tissue versus the combined outcome categories mature teratoma and viable cancer
[2] Mature teratoma versus the combined outcome categories benign tissue and viable cancer
[3] Viable cancer versus the combined outcome categories benign tissue and mature teratoma

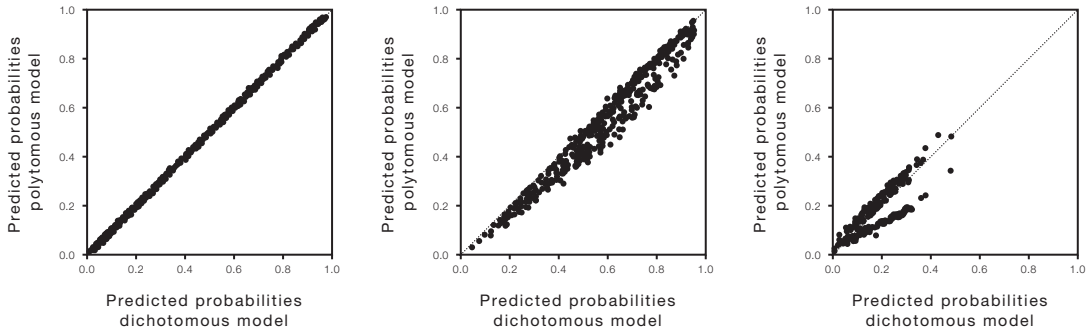
Table 4. Three by three diagnostic classification table for the polytomous (a) and the consecutive dichotomous (b) logistic regression models of the polytomous and consecutive dichotomous models. ROC areas are presented with corresponding 95% confidence intervals.

a

Predicted outcome	Outcome after resection (observed outcome)			
	Benign tissue	Mature teratoma	Viable cancer	Total
Benign tissue	155	28	13	196
Mature teratoma	140	193	30	364
Viable cancer	130	314	91	535
Total	425	535	134	1094

b

Predicted outcome	Outcome after resection (observed outcome)			
	Benign tissue	Mature teratoma	Viable cancer	Total
Benign tissue	132	23	11	166
Mature teratoma	241	300	57	598
Viable cancer	52	212	66	330
Total	425	535	134	1094



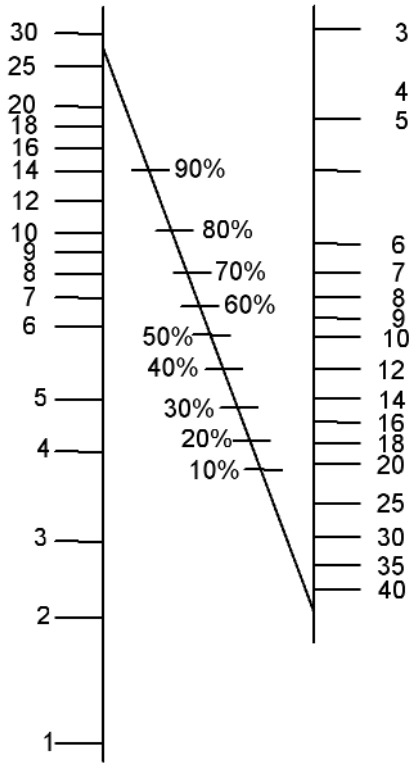
a. Benign tissue b. Mature teratoma c. Viable cancer

Figure 3. **Predicted probabilities estimated by the two consecutive dichotomous models plotted against predicted probabilities estimated by the polytomous model for benign tissue (a), mature teratoma (b) and viable cancer (c). The dashed line indicates perfect agreement between predictions of the polytomous model and the consecutive dichotomous models.**

Figure 4. Score chart for pre-operative estimation of the probability of benign tissue, mature teratoma, and viable cancer in postchemotherapy residual retroperitoneal masses of NSTGCT patients.

Predictor	Score benign tissue		Score mature teratoma	
Primary tumour teratoma negative	4	...	-2	...
Normal AFP serum level	5	...	0	...
Normal HCG serum level	2	...	-2	...
Standardised value of LDH		...		...
0.5	0	...	6	...
0.7	1	...	5	...
1	1	...	4	...
1.5	1	...	3	...
2	2	...	2	...
3	2	...	1	...
4	2	...	0	...
5	2	...	0	...
Postchemotherapy mass size (mm)				...
2	10	...	4	...
5	9	...	4	...
10	8	...	4	...
15	7	...	3	...
20	7	...	3	...
30	5	...	2	...
50	4	...	2	...
70	2	...	1	...
100	0	...	0	...
Reduction in mass size after chemotherapy (%)				...
-50	0	...	3	...
-25	1	...	2	...
0	3	...	2	...
25	5	...	2	...
50	6	...	1	...
75	8	...	1	...
100	10	...	0	...
constant	-11	-11 +	2	2 +
	Sumscore benign tissue	...	Sumscore mature teratoma	...

Total sumscore = sumscore benign tissue + sumscore mature teratoma = ...



The left axis represents the sumscore of benign tissue and the sumscore of mature teratoma, the right axis represents the total sumscore. The diagonal axis represents probabilities. Drawing a line between the left axis and the right axis, one can read the corresponding predicted probability for benign tissue or mature teratoma on the cross-point with the diagonal axis. The probability of viable cancer can be calculated with: 100% - predicted probability (benign tissue) – predicted probability (mature teratoma).

were lower (49% (66/134) and 31% (132/425)) and higher for mature teratoma (56% (300/535)) (Table 4b).

Agreement between the polytomous and dichotomous models

We also plotted predicted probabilities of the polytomous model against the predicted probabilities of the consecutive dichotomous models per outcome category. Figure 3 shows high agreement in predicted probabilities between the polytomous and consecutive dichotomous models for benign tissue, discrepancy was marginal for mature teratoma and more substantial for viable cancer. Notably, for mature teratoma the polytomous model yielded overall lower predicted probabilities, whereas for viable cancer predicted probabilities were either slightly higher or significantly lower as opposed to the consecutive dichotomous models.

Presentation of the polytomous model

Figure 4 shows the polytomous model in a score chart format to facilitate estimation of probabilities for the three outcome categories. This score chart is less straightforward than for dichotomous logistic regression, because the probability of one outcome category depends on the probabilities of the other outcome categories. As an example, we consider a patient with a teratoma negative primary tumour (4 points for benign tissue versus -2 points for mature teratoma), normal prechemotherapy AFP level (5 points versus 0 points), elevated prechemotherapy HCG level (0 points versus 0 points), prechemotherapy standardised LDH value of 3 times normal (2 points versus 1 point), postchemotherapy mass size of 20 mm (7 points versus 3 points), and a reduction in mass size of 50% (6 points versus 1 point). The patient's sumscore is 13 for benign tissue (after subtracting 11 points for the intercept) and 5 for mature teratoma (after adding 2 points for the intercept), and a total sumscore of 18 (13 + 5). Drawing an imaginary line between the sumscore of 13 for benign tissue on the left axis and the total sumscore of 18 on the right axis, one can read the predicted probability for benign tissue on the cross-point with the drawn line, i.e. 55% for this patient. The probability for mature teratoma can be read in a similar way, i.e. 23%. The probability of viable cancer for this patient would be 100% - 55% – 23% = 22%. Based on to the classification rule of 1:3:8, this patient would be classified as having a residual mass with viable cancer.

Discussion

In this article, we examined polytomous logistic regression analysis in diagnostic studies with more than two outcome categories. This method was compared with dichotomous logistic regression analysis. We explained the interpretation of the odds ratios derived from the polytomous model, showed several model performance measures and a user-friendly format (score chart) for application of the polytomous model in daily practice. Using clinically relevant thresholds to classify patients, the polytomous model could better detect viable cancer and benign tissue, compared to the consecutive dichotomous models that could better detect mature teratoma.

Methodological considerations

To appreciate the present results, a few methodological issues should be addressed. First, the most commonly used strategy to model outcomes with more than two categories is to fit separate models for each category. In our example, we could fit a dichotomous model for benign tissue versus viable cancer and another model for mature teratoma versus viable cancer. This would result in similar regression

coefficients as the polytomous model with two main disadvantages. Firstly, the predicted probabilities of the model may sum up to more than 100% and secondly, one overall covariance matrix cannot be estimated, and thus standard errors are incorrect.

Second, when deriving a diagnostic (or prognostic) rule, the power is determined by the number of patients in the smallest group. Usually, this is the category comprising those with the most clinically relevant outcome. For the development of a dichotomous logistic regression model, a rule of thumb is to consider no more than one predictor per ten outcome events, if per predictor one regression coefficient is estimated. The chance of finding spurious associations between predictors and the outcome increases with a decreasing number of events per predictor[28-32]. For the present study, we developed a polytomous logistic regression model with three diagnostic outcome categories meaning that for each predictor two regression coefficients are estimated (formulas 6 and 7). Hence, one should consider no more than one predictor per twenty outcome events. Since the number of outcome events in the smallest category (viable cancer) was 134, the number of predictors was limited to a maximum of six. Actually, we also studied the linearity of the three continuous predictors with restricted cubic splines (with three regression coefficients per spline), thus the maximum number of regression coefficients was exceeded. With increasing outcome categories to be predicted, this limitation deserves even more attention.

Third, before fitting a polytomous regression model potential predictors for each of the outcome categories should be known from previous research. A variable may be a predictor for one or two of the outcome categories but not for the other outcomes. In this study for instance, AFP level was a strong predictor for benign tissue (OR= 2.8), but not for mature teratoma (OR= 0.94). For a polytomous model with very diverse outcome categories, each outcome may require different predictors. This would result in a polytomous model with many predictors and hence many regression coefficients to be estimated. To limit the number of regression coefficients, one can consider some predictors for only one outcome category by setting the coefficients of the other categories to zero[33].

Fourth, the model performance estimated in the data used for model development is usually too optimistic[34-37]. One of the steps in model development therefore is to study and correct for overoptimism with bootstrapping techniques[34;38;39]. Bootstrapping of polytomous models is currently not implemented in statistical software packages. However, it is possible to program bootstrap algorithms in object-oriented packages like S-plus and R. Another step in model development is to shrink the regression coefficients; otherwise, predictions will be too extreme for new patients. A heuristic shrinkage factor can be estimated like for dichotomous logistic regression[40]:
(model X^2 – number of regression coefficients) / model X^2

Finally, for predicting several unordered outcome categories, Wijesinha et al and Begg et al preferred a series of consecutive dichotomous regression models to approximate one overall polytomous model, since in 1983 polytomous regression analysis was not yet implemented in standard statistical software and computer storage was limited[6;7]. These practical limitations have now been overcome and thus are no longer an issue.

Clinical implications

The overall performance of the polytomous model was slightly better than the consecutive dichotomous models. The polytomous model particularly discriminated better viable cancer (ROC areas 0.66 and 0.64). This can be explained by the modelling procedures. With polytomous logistic regression analysis, viable cancer is modelled completely separate from mature teratoma. The regression coefficients are estimated for benign tissue versus viable cancer and for mature teratoma versus viable cancer. With consecutive dichotomous logistic regression analysis, viable cancer is separated from mature teratoma only in the second model, i.e. regression coefficients are also estimated for mature teratoma versus viable cancer, but benign tissue is compared to the combination of mature teratoma and viable cancer. Since in our study the category mature teratoma included much more patients than the viable cancer category, this model mainly discriminated between benign tissue and mature teratoma. Hence, the proportion of detected mature teratoma case was higher as opposed to the polytomous model at the expense of a lower proportion of detected patients with benign tissue and viable cancer cases.

From a clinical point of view, it is reasonable to use a high threshold value to classify masses as benign tissue or mature teratoma. This results in a low threshold value for viable cancer, which implies that most masses with viable cancer will be correctly classified. Hence, the most serious outcome will be detected and surgically resected as much as possible. The dilemma lies in weighing the risks and costs of unnecessary resection of benign tissue, against the risks (e.g. relapse and impaired survival) and costs of missing mature teratoma or viable cancer. With the three by three diagnostic classification table of the polytomous regression model, 36% of the benign tissue and 36% of mature teratoma cases were detected with a rather high proportion of detected viable cancer cases (68%). For the consecutive dichotomous models, the proportion of detected mature teratoma case was higher (56%) at the expense of a lower proportion of detected patients with benign tissue and (31%) and viable cancer cases (49%). Accordingly, the polytomous model showed better performance in detecting viable cancer cases than the consecutive dichotomous models at the proposed threshold values.

The discrimination between viable cancer and benign tissue is of predominant importance for the decision to resect a residual mass and the discrimination between mature teratoma and the other outcomes is of second consideration. Hence, the benefit of discriminating mature teratoma from viable cancer with our polytomous model may be questionable. However, the advantage of discriminating mature teratoma from viable cancer preoperatively may be helpful for decision making with taking patients' personal preferences into account (e.g. the consideration to postpone resection when extensive surgery may damage adjacent structures).

Compared to a dichotomous model that predicts the presence or absence of a particular disease (formula 2), the polytomous model might seem more complex for application in clinical practice (formula 6 and 7). However, with the presented score chart we have tried to overcome this problem, since patient probabilities for the three outcome categories can easily be estimated.

In conclusion, our analyses point to a valuable role for the polytomous logistic regression model in the prediction of several diagnostic outcome categories. Simultaneous prediction of several diagnostic outcome probabilities particularly applies to diagnostic situations in which commonly various potential diagnoses are considered in a patient presenting with particular signs and symptoms. Therefore, the method of polytomous regression analysis may serve clinical practice better than conventional dichotomous regression analysis, and deserves closer attention in future diagnostic research.

chapter

8

Concluding remarks

Where are we now?

In the era of evidence-based medicine, diagnostic procedures also need to undergo critical evaluations. In contrast to guidelines for randomised trials and observational etiologic studies, principles and methods for diagnostic evaluations are still incomplete[1;2]. The research described in this thesis was conducted to further improve the methods for design and analysis of diagnostic studies.

Design issues

In the past, most diagnostic accuracy studies followed a univariable or single test approach with the aim to quantify the sensitivity, specificity or likelihood ratio. However, single test studies and measures do not reflect a test's added value. It is not the singular association between a particular test result or predictor and the diagnostic outcome that is informative, but the test's value independent of diagnostic information. Multivariable modelling is necessary to estimate the value of a particular test conditional on other test results. However, diagnostic prediction rules are not the solution to everything. They have certain drawbacks, such as overoptimistic accuracy when applied to new patients. Recently, methods have been described to overcome some of these drawbacks[3-6].

Typically, in diagnostic research one selects a cohort of patients with an indication for the diagnostic procedure at interest as defined by the patients' suspicion of having the disease of interest. The data are analysed cross-sectionally. When appropriate analyses are applied, results from nested case-control studies should be virtually identical to results based on a full cohort analysis. We showed that the nested case-control design offers investigators a valid and efficient alternative for a full cohort approach in diagnostic research. This may be particularly important when the results of the test under study are costly or difficult to collect.

It is suggested that randomised controlled trials deliver the highest level of evidence to answer research questions[7-9]. The paradigm of a randomised study design has also been applied to diagnostic research[10-14]. We described that a randomised study design is not always necessary to evaluate the value of a diagnostic test to change patient outcome. A test's effect on patient outcome can be inferred and indeed considered as quantified -using decision analysis- 1) if the test is meant to include or exclude a disease for which an established reference is available, 2) if a cross-sectional accuracy study has shown the test's ability to adequately detect the presence or absence of that disease based on the reference, and finally 3) if proper, randomised therapeutic studies have provided evidence on efficacy of the optimal management of this disease[10;11;15;16]. In such instances diagnostic research does not require an additional randomised comparison between two (or more) 'test-treatment strategies' (one with and one without the test under study) to establish the test's effect on patient outcome. Accordingly, diagnostic research -including the quantification of the effects of diagnostic testing on patient outcome- may be executed more efficiently.

1_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

2_ Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7-18.

3_ Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1986;1:54-77.

4_ Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros. Stat Med 1996;15:361-87.

5_ Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54: 774-81.

6_ Steyerberg EW, Borsboom GJ, Van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567-86.

7_ Trout KS. How to read clinical journals: IV. To determine etiology or causation. Can Med Assoc J 1981;124:985-90.

8_ Sacks H, Chalmers TC, Smith H.
Randomized versus historical controls for clinical trials.
Am J Med 1982;72:233-40.

9_ Pocock SJ, Elbourne DR.
Randomized trials or observational tribulations?
N Engl J Med 2000;342: 1907-9.

10_ Dixon AK.
Evidence-based diagnostic radiology.
Lancet 1997;350:509-12.

11_ Fryback D, Thornbury J.
The efficacy of diagnostic imaging.
Med Decis Making 1991;11:88-94.

12_ Hunink MG, Krestin GP.
Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology.
Radiology 2002;222:604-14.

13_ Mackenzie R, Dixon AK.
Measuring the effects of imaging: an evaluative framework.
Clin Radiol 1995;50:513-18.

14_ Bossuyt PP, Lijmer JG, Mol BW.
Randomised comparisons of medical tests: sometimes invalid, not always efficient.
Lancet 2000;356:1844-7.

15_ Moons KG, Grobbee DE.
Diagnostic studies as multivariable, prediction research.
J Epidemiol Community Health 2002;56:337-8.

16_ Moons KG, Ackerstaff RGA, Moll FL, Spencer MP, Algra A.
Association of intraoperative transcranial doppler monitoring variables with stroke from carotid endarterectomy (respons).
Stroke 2001;32:813.

Analytical issues

Diagnostic research aims to quantify a test's added contribution given other diagnostic information available to the physician in determining the presence or absence of a particular disease. Commonly, diagnostic prediction rules use dichotomous logistic regression analysis to predict the presence or absence of a disease. We showed that genetic programming and polytomous modelling are promising alternatives for the conventional dichotomous logistic regression analysis to develop diagnostic prediction rules. The main advantage of genetic programming is the possibility to create more flexible models with better discrimination. This is especially important in large data sets in which complex interactions between predictors and outcomes may be present.

Using polytomous logistic regression, one can directly model diagnostic test results in relation to several diagnostic outcome categories. Simultaneous prediction of several diagnostic outcome probabilities particularly applies to situations in which more than two disorders are considered in the differential diagnoses. As this is commonly the case, polytomous regression analysis may serve clinical practice better than conventional dichotomous regression analysis. Both alternatives deserve closer attention in future diagnostic research.

We also showed that the development of a diagnostic prediction rule is not the end of the 'research line', even when a rule is subsequently adjusted for optimism using internal validation techniques e.g. bootstrap techniques. External validation of such rules in new patients is always required before introducing a rule in daily practice. This indicates that internal validation of prediction models may not be sufficient and indicative for the model's performance in future patients. Rather than viewing a validation data set as a separate study to estimate an existing rule's performance, validation data may be combined with data of previous derivation studies to generate more robust prediction models using recently suggested methods[6;17-20].

Perspectives for future diagnostic research

First, genetic programming and polytomous regression analysis require additional evaluation before they can be considered a standard approach to develop multivariable diagnostic rules.

Second, many (prognostic and diagnostic) prediction rules have been developed and there are numerous examples of rules showing lower predictive accuracy in new patients[17;21-24]. However, when validating or testing a developed prediction rule in new patients, it is largely unknown which factors truly compromise the accuracy of the rule in these new patients[25;26]. We believe that the methodology for external validation studies that aim to test the generalisability of diagnostic rules in new patients should be improved and refined.

Decrease in predictive accuracy in a validation study with new patients does not automatically imply that a prediction rule was inadequately developed. It could be due to differences between the derivation and validation population such as differences in outcome frequency, in case mix (i.e. distribution of predictors), and in strength of associations between test results and disease. For instance, to discriminate patients with and without the disease in a validation set with a more homogeneous case mix is more difficult than with a heterogeneous case mix.

The generalisability of a prediction rule may also be affected when an important test result was not included in the rule during model development, or when the rule includes one or more spurious predictors. Some of these issues have been addressed in the statistical literature[6;18;27], but their impact, either alone or in combination, have hardly been studied on empirical data. Therefore, the mechanisms that lead to a reduced generalisability of prediction rules in new patients deserve particular attention in future research.

Finally, improved methods for design and analysis of external validation studies of diagnostic prediction rules do not guarantee good applicability or improved quality of care. Applicability of a rule is also determined by the way it is presented and how it can be used in practice to estimate disease probabilities in individual patients[28;29]. To improve quality of care rules must not only be adequately developed and validated, they must also be used by doctors[30]. Accordingly, after being developed, internally and externally validated, studies are needed to investigate whether patient outcome indeed is improved when using a prediction rule. If so, then widespread implementation of the rule is indicated[26;31].

17_ Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE et al.
External validation is necessary in prediction research: A clinical example.
J Clin Epidemiol 2003;56:826-32.

18_ Houwelingen van JC.
Validation, calibration, revision and combination of prognostic survival models.
Stat Med 2000;19:3401-15.

19_ Houwelingen van JC.
Shrinkage and penalized likelihood methods to improve diagnostic accuracy.
Stat Neerl 2001;55:17-34.

20_ Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG.
Internal and external validation of predictive models: a simulation study of bias and precision in small samples.
J Clin Epidemiol 2003;56:441-7.

21_ Kneyber MC, Moons KG, Groot RD, Moll HA.
Prediction of duration of hospitalization in respiratory syncytial virus infection.
Pediatr Pulmonol 2002;33:453-7.

22_ Sanson B, Lijmer JG, Mac Gillavry MR, Turkstra F, Prins MH, Buller HR.
Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism.
Thromb Haemost 2000;83:199-203.

23_ Lim WS, Lewis S, Macfarlane JT.
Severity prediction rules in community acquired pneumonia: a validation study.
Thorax 2000;55:219-23.

24_ Fortescue EB, Kahn K, Bates DW.
Prediction rules for complications in coronary bypass surgery: a comparison and methodological critique.
Med Care 2000;38:820-35.

25_ Altman DG, Royston P.
What do we mean by validating a prognostic model?
Stat Med 2000;19:453-73.

26_ Justice AC, Covinsky KE, Berlin JA.
Assessing the generalizability of prognostic information.
Ann Intern Med 1999;130: 515-24.

27_ Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S.
A comparison of goodness-of-fit tests for the logistic regression model.
Stat Med 1997;16:965-80.

28_ Feinstein AR.
"Clinical Judgment" revisited: the distraction of quantitative models.
Ann Intern Med 1994;120: 799-805.

29_ Puhan MA, Steurer J, Bachmann LM, ter Riet G.
Variability in diagnostic probability estimates.
Ann Intern Med 2004;141:578-9.

30_ Reid MC, Lane DA, Feinstein AR.
Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy.
Am J Med 1998;104:374-80.

31_ Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L et al.
Effectiveness and efficiency of guideline dissemination and implementation strategies.
Health Technol Assess 2004;8:iii-72.

Summary

To set a diagnosis is the cornerstone for medical care as it indicates treatment and provides indirectly an estimate of the patient's prognosis. A diagnostic test commonly has no direct therapeutic effects and does not directly influence patient outcome; setting a diagnosis is rather a vehicle to guide therapies. The motive for diagnostic research is commonly efficiency: decreasing the patient burden and costs of the diagnostic work-up in practice, considering the consequences of false diagnoses. Various reviews have demonstrated that the majority of published diagnostic accuracy studies still have methodological flaws in design or analysis or provide results with limited practical applicability. This has been attributed to the absence of proper principles and methods for diagnostic research. The research described in this thesis was conducted to further improve the methods for design and analysis of diagnostic studies.

In chapter 2 we argue why diagnostic studies rather follow a multivariable approach to assess the added value of a diagnostic test than a single test approach. In our view, the difference between test research and diagnostic research has received too little attention in the majority of articles concerning principles and methods for diagnostic research. With test research we refer to studies that follow a single test or univariable approach. With diagnostic research we refer to studies that aim to quantify a test's added contribution beyond test results readily available to the physician, in the estimation of presence of a particular disease. We believe that test research has limited applicability to clinical practice and we explain why this is the case, followed by a brief description of the multivariable approach, and two clinical examples illustrating the hazards of test research. Finally, we describe the few situations in which test research may be worthwhile, i.e. in the context of screening and in the initial phase of developing a new test.

The message of chapter 3 is that a randomised study design in diagnostic research is often not necessary to quantify whether a new test may lead to improved patient outcome. In almost every system that grades epidemiological studies according to their level of evidence, randomised studies or meta-analyses of randomised studies receive the highest classification. Accordingly, it has widely been advocated that after establishing a test's diagnostic accuracy, the impact of the test on patient outcome must also be quantified. To do so, the use of randomised comparisons has been proposed, since randomisation has become the general paradigm to study the effect of interventions on patient outcome. However, to validly demonstrate the beneficial effect of a diagnostic procedure on patient outcome, we believe that randomisation is not by definition a prerequisite. In many situations, randomised studies in diagnostic research are not necessary and cross-sectional accuracy studies are fully acceptable to validly estimate the value of the diagnostic test in improvements of patient care.

Chapter 4 investigates the use of a nested case-control approach in diagnostic research. Diagnostic studies that aim to quantify the value of tests, commonly apply a cross-sectional cohort design. Also by the recent STARD guideline, use of a conventional case-control design has been disapproved for diagnostic research, because application of this design often results in biased estimates of diagnostic accuracy measures. However, when appropriate analyses are applied, results from nested case-control studies should be virtually identical to results based on

a full cohort analysis. To illustrate this, we used empirical data comprising patients suspected of deep vein thrombosis (DVT). Nested case-control data samples were drawn from the full cohort to compare estimated diagnostic accuracy measures for two predictors (D-dimer test and calf difference test) associated with DVT with those obtained from full cohort analysis. Diagnostic accuracy measures of the D-dimer test and calf difference test estimated in the nested case-control samples were very similar to those in the full cohort. Especially when predictor variables are costly or difficult to collect, a nested case-control design reduces the number of subjects for whom test results are needed compared to the full cohort approach. Therefore, the STARD guideline should be updated by stating that the nested case-control approach is a valid and efficient design for diagnostic research.

In chapter 5, we validated a previously derived multivariable prediction rule for neurological sequelae after childhood bacterial meningitis in a large sample of children with bacterial meningitis selected from almost all hospitals in The Netherlands. The rule showed very poor agreement between predicted and observed risks in the calibration plot and by the Hosmer-Lemeshow test (p-value < 0.01). The ROC area was 0.65 (95%CI: 0.57-0.72), which was statistically significant lower than the ROC area in the derivation set (0.87 (0.78-0.96), p-value < 0.01). This indicated that internal validation of prediction models by bootstrap techniques may not be sufficient and indicative for the model's performance in future patients. We therefore updated the original rule by re-estimating the regression coefficients of the original predictors and added extra predictors, after combining the data of the derivation and validation set. In the combined data set, gender was no longer a predictor. The updated rule included two additional predictors, and showed better performance than the original rule. After adjustment for optimism, the ROC area was 0.77 (95%CI: 0.72-0.82). Our analyses again demonstrate the importance of using new (validation) data to test existing prediction rules. Rather than viewing a validation data set as a separate study to estimate an existing rule's performance, validation data can often better be combined with data of previous derivation studies to generate more robust prediction models.

Chapter 6 describes a study to evaluate the value of genetic programming as compared to the well-known and widely applied multivariable logistic regression analysis for diagnostic research questions. Genetic programming is a search method that can be used to solve complex associations between large numbers of predictor variables. We used empirical data from patients suspected of pulmonary embolism (PE). Using part of the data (67%), we developed and internally validated a diagnostic prediction model by genetic programming and by logistic regression, and compared their predictive accuracy in the remaining data (validation set). In the validation set, the area under the ROC curve of the genetic programming model was larger (0.73; 95%CI: 0.64-0.82) than that of the logistic regression model (0.68; 0.59-0.77). The calibration of both models was the same, indicating a similar amount of overoptimism. Although the interpretation of a genetic programming model is less intuitive, genetic programming seems a promising technique to develop prediction rules for diagnostic and prognostic purposes, in particular when the aim is to achieve optimal discrimination.

Chapter 7 examines the value of polytomous regression in diagnostic studies. Polytomous logistic regression analysis has been advocated when there are more than two unordered outcome categories. We used empirical data from a study on

diagnosing residual retroperitoneal mass histology in patients with nonseminomatous testicular germ cell tumour. The differential diagnoses in these patients included benign tissue, mature teratoma and viable cancer. The performance and use of a polytomous model to estimate the probability of each of the three diagnoses was compared with that of two consecutive dichotomous logistic models. The ROC areas for benign tissue (both 0.83), mature teratoma (both 0.78) and viable cancer (0.66 and 0.64) were almost the same for both models. Also the calibration, R^2 and Brier score were very similar for both models. However, using clinically relevant thresholds, 68% of the viable cancer cases were detected with polytomous regression analysis, whereas only 49% of the viable cancer cases were detected with the two consecutive dichotomous regression models. We concluded that when several diagnostic outcome categories are initially considered, polytomous regression modelling may serve clinical practice better than conventional dichotomous modelling and deserves close attention in future diagnostic research.

In Chapter 8 we make some concluding remarks on the findings described in this thesis and we outline perspectives for future research.

Samenvatting

Diagnostiek ligt aan de basis van elke medische behandeling en verschaft indirect een prognose voor de patiënt. Een diagnostische test heeft gewoonlijk geen therapeutisch effect en heeft daarom geen directe invloed op gezondheidswinst voor de patiënt. Het kan een arts wel helpen om een juiste therapie te kiezen. Omdat elke vorm van diagnostiek belastend is voor de patiënt, en tijd en geld kost, moet het diagnostische proces zo efficiënt mogelijk worden uitgevoerd met inachtneming van acceptabele percentages verkeerde beslissingen. Uit verscheidene review artikelen is gebleken dat een meerderheid aan diagnostische studies nog steeds methodologische gebreken vertoont wat betreft hun studie opzet en methode van analyseren. Hierdoor zijn resultaten van deze studies vaak van beperkte waarde voor de medische praktijk. Dit alles kan in belangrijke mate worden toegeschreven aan het ontbreken van een methodologisch protocol voor de opzet en uitvoering van diagnostisch onderzoek, zoals dat al wel bestaat voor etiologisch en interventie onderzoek. In dit proefschrift worden methoden beschreven voor verbetering van studie opzet en analyse in diagnostisch onderzoek.

In hoofdstuk 2 beargumenteren we waarom het multivariabele en hiërarchisch gefaseerde karakter van het diagnostische proces weerspiegeld zou moeten worden in de vraagstelling, studie opzet, analyse en presentatie van diagnostisch onderzoek. Met univariabele test evaluaties bedoelen we studies die zich richten op het schatten van de sensitiviteit, specificiteit, likelihood ratio of discriminatie, gemeten met het oppervlak onder de 'receiver operating characteristic' (ROC) curve, van één bepaalde test. We noemen dit test evaluatie onderzoek, omdat het slechts testkarakteristieken schat. Met diagnostisch onderzoek bedoelen we studies die de toegevoegde waarde van een test voor het schatten van de aan- of afwezigheid van een bepaalde ziekte kwantificeren, gegeven de kennis die de arts al ter beschikking heeft uit eerdere fasen van het diagnostisch proces, zoals de anamnese en het lichamelijk onderzoek. Aan de hand van voorbeelden lichten we toe dat test evaluatie onderzoek slechts beperkte waarde heeft voor toepassing in de praktijk. Verder geven we een korte beschrijving van een ons inziens meer correcte benadering van diagnostisch onderzoek, namelijk de multivariabele methode.

De belangrijkste boodschap van hoofdstuk 3 is dat een gerandomiseerde studie opzet niet altijd vereist is om het effect van een nieuwe test op gezondheidswinst voor de patiënt op een valide wijze te kwantificeren. Gerandomiseerde studies (of meta-analyses van gerandomiseerde studies) worden vaak gezien als de beste studies voor het uitvoeren van epidemiologisch onderzoek. Daarom wordt aanbevolen dat wanneer de toegevoegde waarde van een diagnostische test wordt onderzocht, ook het effect van deze test op de gezondheidswinst voor de patiënt moet worden gekwantificeerd met een gerandomiseerde studie opzet. Wij beschrijven echter de redenen waarom voor diagnostisch onderzoek een cross-sectionele studie vaak voldoende is en een gerandomiseerde studie overbodig is voor het bepalen van de toegevoegde waarde van een test.

In hoofdstuk 4 wordt het gebruik van een geneste patiënt-controle studie onderzocht voor diagnostisch onderzoek. Voor studies naar de diagnostische (toegevoegde) waarde van een test wordt vaak een cross-sectionele cohort studie opzet toegepast. In ondermeer de recent geformuleerde STARD richtlijn voor het valide uitvoeren

van diagnostisch onderzoek, wordt het gebruik van een patiënt-controle studie opzet afgeraden, omdat dit vaak zou leiden tot onzuiver geschatte uitkomstmaten. Echter, wanneer valide analyse technieken worden toegepast, zijn de resultaten van een geneste patiënt-controle studie vrijwel identiek aan resultaten die gebaseerd zijn op het hele cohort. Dit hebben we geïllustreerd aan de hand van een data set (cohort) met patiënten die verdacht werden van diep veneuze trombose (DVT). Geneste patiënt-controle samples werden getrokken uit het hele cohort om de geschatte diagnostische uitkomstmaten van twee predictoren (de D-dimer test en het verschil in omvang tussen het been met en zonder trombose) te vergelijken met de uitkomstmaten van het hele cohort. De diagnostische uitkomstmaten van de geneste patiënt-controle samples kwamen overeen met de uitkomstmaten van het hele cohort. Vooral als het verzamelen van bepaalde variabelen moeilijk is of kostbaar, is de toepassing van een geneste patiënt-controle studie een efficiënte methode. De STARD richtlijn zou moeten worden aangepast met de opmerking dat de geneste patiënt-controle studie een valide en efficiënte studie opzet is voor diagnostisch onderzoek.

In hoofdstuk 5 wordt beschreven hoe we een eerder ontwikkelde klinische voorspelregel voor neurologische restverschijnselen na bacteriële meningitis bij kinderen hebben bestudeerd in nieuwe patiënten (validatie) en aangepast aan de hand van deze nieuwe data. De voorspelregel vertoonde een slechte overeenkomst (calibratie) tussen de werkelijke aanwezigheid van neurologische restverschijnselen en de door de regel voorspelde kans hierop. Het onderscheidend vermogen (discriminatie), uitgedrukt in het oppervlak onder de ROC curve was 0.65 (95% BI: 0.57-0.72), en dat was statistisch significant lager dan in de derivatie set (0.87 (0.78-0.96), p< 0.01). Dit duidt erop dat het intern valideren van voorspelregels met ‘bootstrappen’ niet altijd voldoende is om realistische schattingen van de model prestaties te krijgen in nieuwe patiënten. Daarom hebben we de originele voorspelregel aangepast door de regressie coëfficiënten van de predictoren (voorspellende variabelen) in de voorspelregel opnieuw te schatten en nieuwe predictoren toe te voegen op basis van de gecombineerde data van de ontwikkelings- en validatie studie. Geslacht was niet langer een predictor in deze gecombineerde data set. Na toevoeging van twee extra predictoren was het oppervlak onder de ROC curve 0.77 (95%BI: 0.72-0.82). Validatie data kunnen dus niet alleen gebruikt worden voor het testen van een bestaande voorspelregel, maar ook worden gecombineerd met data van de voorafgaande ontwikkelingsstudie, om meer betrouwbare kansschattingen te verkrijgen.

Hoofdstuk 6 vergelijkt ‘genetic programming’, een methode om complexe relaties tussen variabelen en de uitkomst te kwantificeren, met de bekende multivariabele logistische regressie analyse voor diagnostisch onderzoek. Hiervoor werden data gebruikt van patiënten met klachten en symptomen die kunnen wijzen op een longembolie. Met tweederde deel van deze data set werd een voorspelregel ontwikkeld met genetic programming en een voorspelregel met multivariabele logistische regressie analyse. De voorspellende waarden van de twee regels werd bepaald en onderling vergeleken in de rest (eenderde) van de data set. In deze validatie data was het oppervlak onder de ROC curve van de genetic programming regel groter (0.73; 95%BI: 0.64-0.82) dan die van de regel verkregen met logistische regressie (0.68; 0.59-0.77). De calibratie was gelijk voor beide methoden. Genetic programming levert lastig te interpreteren resultaten op, maar is toch een veelbelovende methode om voorspelregels te maken voor diagnostische en prognostische doeleinden.

In hoofstuk 7 wordt de waarde onderzocht van polytome logistische regressie analyse voor diagnostisch onderzoek. Polytome regressie lijkt vooral voordelen te bieden als er meer dan twee ongeordende uitkomstcategorieën zijn. We maakten gebruik van empirische data van een studie naar het voorspellen van de histologie van restweefsel bij patiënten die behandeld waren voor metastasen van een nonseminomateuze kiemcel tumor. De differentiaal diagnose van de histologie van het restweefsel bij deze patiënten bestaat uit benigne weefsel, matuur teratoom en maligne weefsel. De prestaties van de polytome voorspelregel werden vergeleken met de prestaties van twee opeenvolgende dichotome voorspelregels. De oppervlakten onder de drie ROC curves waren bijna gelijk voor beide modellen: 0.83 voor benigne weefsel, 0.78 voor matuur teratoom en 0.66 (polytome regel) en 0.64 (dichotome regel) voor maligne weefsel. Ook de calibratie, R^2 en Brier score waren vergelijkbaar. Echter bij het hanteren van klinisch relevante grenswaarden voor het classificeren van patiënten, werd 68% van de patiënten met maligne weefsel gedetecteerd met polytome regressie, terwijl slechts 49% van de patiënten met maligne weefsel gedetecteerd werd met dichotome regressie analyse. We concluderen dat polytome regressie een betere methode is bij het voorspellen van meerde diagnostische uitkomsten dan dichotome regressie analyse. Deze methode heeft daarom onze voorkeur voor het ondersteunen van besluitvorming in de medische praktijk.

Tenslotte maken we in hoofdstuk 8 enkele concluderende opmerkingen over onze bevindingen en geven we suggesties voor toekomstig onderzoek.

Dankwoord

Om te beginnen wil ik elke lezer bedanken met ‘thank you for just being who you are’. Onderstaande personen wil ik in het bijzonder noemen.

Prof. Dr. Grobbee, beste Rick, hartelijk dank voor de inspirerende opmerkingen en aanvullingen op de artikelen en het manuscript als geheel. Je wist altijd de juiste snaar te raken om me aan het denken te zetten.

Prof. Dr. K.G.M. Moons, Carl, dankzij jouw begeleiding heb ik me op een heel goede manier kunnen ontwikkelen. Meer dan wie ook heb jij me gestimuleerd voor het onderzoek beschreven in dit proefschrift. Je wist me, ook bij tegenslag, op het juiste spoor te houden. Ik ben er trots op dat je mijn promotor bent.

De overige leden van het ‘predictieclubje’: Yvonne Vergouwe, Jolanda van den Bosch, en Kristel Janssen. We hebben veel van elkaar geleerd tijdens het wekelijks overleg. Yvonne, veel dank voor het meedenken en je begeleidende rol in het laatste jaar van mijn aanstelling. Je invloed is groter geweest dan je zelf denkt.

Alle co-auteurs bedankt voor jullie bijdrage aan de artikels.

Alle kamergenoten met wie ik een goede tijd heb gehad op het Julius Centrum en alle overige collega’s: het ga jullie goed in de toekomst.

Paranymfen (Machiel en Kristel) en alle vrienden, hartelijk dank voor jullie vriendschap.

(Schoon)familie: het is goed om zo’n prettige band met elkaar te hebben.

Pa, ma, Ella en Peter: ik kan me geen betere ouders, zus en broer indenken dan jullie.

Natasja, sinds wij elkaar ontmoet hebben is het leven nog veel mooier geworden. Met jou ga ik het leven delen!

Curriculum Vitae

Corné Biesheuvel was born on April 3rd, 1976 in Brakel, The Netherlands. In 1995, after graduating secondary school at Oude Hoven in Gorinchem, he started his training in Medical Biology at the Faculty of Medicine, of the Utrecht University. For a research project on hepatitis B vaccination he went to Batam, Indonesia. As part of his training he conducted a research project at the Laboratory for Experimental Neurology of the UMC Utrecht on a transgenic mouse model for amyotrophic lateral sclerosis. Next, he performed a research project on constipation in the elderly at the Department of Medical Physiology and Sports Medicine of the UMC Utrecht. He was awarded with an incentive prize for young researchers for the report of this project. He graduated in August, 2000. In August 2001, he started the studies described in this thesis at the Julius Center for Health Sciences and Primary Care of the UMC Utrecht (supervised by Prof. Dr. D.E. Grobbee and Prof. Dr. K.G.M. Moons). He obtained his Master of Science in Clinical Epidemiology at the Netherlands Institute for Health Sciences (NIHES), Erasmus MC, Rotterdam in June 2003. Currently, he is working as a postdoctoral researcher at the Department of Radiology of the UMC Utrecht.