Diagnostic and prognostic models: applications and methods

Peter Zuithoff

ISBN: 978-94-6108-284-8
Lay-out: Nicole Nijhuis - Gildeprint Drukkerijen
Printed by: Gildeprint Drukkerijen
Printed on FSC certified paper

Diagnostic and prognostic models: applications and methods

Diagnostische en prognostische modellen: toepassingen en methodologie
(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. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 17 april 2012 des middags te 12.45 uur

door

Nicolaas Pieter Adriaan Zuithoff

geboren op 22 juni 1967 te Rotterdam

Promotor: Prof.dr. K.G.M. Moons

Co-promotoren: Dr. M.I. Geerlings
Dr. Y. Vergouwe

The studies presented in this thesis were financially supported by the following: the European Commission (ref. PREDICT-QL4-CT2002-00683), the Netherlands Organization for Scientific Research (ref. ZonMw 016.046.360, project 9120.8004 and project 918.10.615), educational grants of all cardiac electronic device providers in the Netherlands, Biotronic Nederland BV, Boston Scientific BV, Medtronic Netherlands BV, Sorin/ELA BV, St Jude Medical BV, The Dutch Pacemaker Registry (SPRN), Groningen, and the Heart and Lung foundation, Utrecht, The Netherlands.

Contents

Chapter 1: General Introduction 7

Part I: Clinical applications

Chapter 2: A clinical prediction rule for deteting major depressive disorder in primary care: the PREDICT-NL study 17

Chapter 3: The Patients Health Questionnaire-9 for detecting major depressive disorder in primary care: consequences of current thresholds in a crosssectional study 35

Chapter 4: Predicting postherpetic neuralgia in elderly primary care patients with herpes zoster: prospective prognostic study 49

Chapter 5: Incidence and predictors of short and long term complications in pacemaker therapy: the FOLLOWPACE study 63

Part II: Methodological issues

Chapter 6: Reporting of aims, designs, predictors, and outcomes in clinical prediction research: a systematic review 85

Chapter 7: Reporting of statistical methods, predictive performance and validation techniques in clinical prediction research: a systematic review 107

Chapter 8: Principal Components Analysis to reduce the number of candidate predictors in diagnostic and prognostic prediction modeling 125

Chapter 9: Dichotomising continuous outcomes in prediction research, a bad idea? 147

Chapter 10: General Discussion 165

Summary 173
Nederlandse samenvatting 179
Dankwoord 187
CV 195

Introduction

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

1

Prediction modelling, both diagnostic and prognostic, has become a major topic in clinical research and practice[1–10]. Traditionally, clinicians intuitively combine and judge the documented patient information, on e.g. risk factors and test results, to implicitly assess the probability or risk of having (in diagnostic estimations) or developing (in prognostic estimations) for certain diseases or outcomes. Nowadays, prediction models become increasingly available to more explicitly assess these disease and outcome probabilities more formally and objectively in their decision making, and results are increasingly incorporated in guidelines for disease risk and management. Clinical prediction models can thus be used to predict diagnostic and prognostic outcomes[3,11–16]. Diagnostic models combine patient characteristics and test results to predict the presence of a certain disease in suspected patients. Prognostic models combine predictors to predict the future occurrence of a certain outcome in patients at risk for those outcomes. Accurate prediction models are valuable to inform patients (e.g. to encourage lifestyle changes if appropriate), to aid clinical decisions, and to stratify patients for more efficient designs of randomised therapeutic trials. The broad interest in prediction models, require accurate methods for model development[1–4,16–21]. Poor research methodology can lead to poor performing prediction models. Various authors have provided guidelines and recommendations on the design and statistical analysis of prediction modelling[2,5,11,12,14,17,18,22–29].

Briefly, from a methodological perspective, prediction models are ideally constructed on sufficiently large numbers of participants, included in a study with an adequate design[1,3,5,17,18]. Candidate predictors for the models are ideally selected from theoretical or clinical understanding of the outcome to be predicted and from previously conducted prediction studies[11,17,18]. If necessary, strategies to properly deal with missing values should be performed[25,30–32]. Non-linear associations between predictors and the outcome should be assessed and appropriately incorporated in the model development, preferably derived from theoretical or clinical knowledge as well[17,18]. The best combination of predictors are ideally selected for inclusion in the final model. The performance of the model, i.e. the ability of the model to estimate an accurate risk of the outcome of interest (calibration) as well as the ability to distinguish between patients with and without the outcome of interest (discrimination), should ideally be tested in new participants that were not used for the model development (external validation).

From a clinical perspective, models should include as few predictors as possible and as easy to document or measure as possible, while at the same time offering sufficient predictive ability. However, prior knowledge is frequently insufficient to select the best predictors in advance. Predictor selection is therefore regularly based on the same data that are used for model development, which increases the risk of developing models too much fitted to the data at hand, reducing their applicability in and generalizability to new individuals

Clearly, the aim for developing clinically useful prediction models, including considerations on the burden for patients, physicians and health care budgets, may interfere with methodological recommendations and guidelines. For example, the inclusion of sufficiently large number of participants in clinical studies is often limited by practical constraints. Studies on outcomes with a low occurrence rate, would require the inclusion of excessive numbers of participants. Also, the frequent use of predictor selection techniques during model development frequently means that too many candidate predictors are studied in the first place in relation to the number of available study participants. Predictor selection from a large set of candidate predictors in small sample sizes may introduce instability in prediction models, which will threaten the validity of predictions in new participants[11,17,29,33,34].

This thesis has 2 parts. In part 1 we developed various clinical prediction models, both diagnostic and prognostic, in different disease areas, using empirical data sets of e.g. limited size, starting with relatively large numbers of candidate predictors, including various non-linear relationships between predictors and outcomes, and using different measures of predictive performance. The second part first addresses the current reporting and methodological conduct in clinical prediction research. Further, we explore alternative statistical methods for prediction research with limited sample sizes, large numbers of candidate predictors, and dealing with linear versus dichotomous outcomes .

Outline of this thesis

Part I thus describes various empirical prognostic and diagnostic model development and validation studies addressing a broad scope of issues commonly encountered prediction modelling. Chapter 2 develops and validates a diagnostic prediction model for the detection of major depressive disorders in primary care patients. The prediction model was specifically designed to aid general practitioners in daily practice, considering the burden and social consequences of this disorder. We also externally validated a simple, easy-to-use questionnaire for detection of major depressive disorder, the PHQ-9 that may be used to screen for major depressive disorder. The results of this validation study are described in chapter 3. In chapter 4, prediction modelling is used to identify the best prognostic predictors for classifying the herpes zoster patients at highest risk for the development of postherpetic neuralgia, with a view to applicability in primary care. In chapter 5, prognostic predictors for short term and long term complications in patients with a pacemaker are studied.

Part II addresses studies on the methodological issues related to prediction studies. Chapters 6 and 7 present a systematic review of the reporting and methodology used in prediction studies in the current literature. We assess to what extent prediction research was reported and conducted according to current methodological guidelines. Chapter 6 focuses on the reporting

of aim, design, study samples, and definition of outcome and predictors; chapter 7 focuses on the reporting and conduct of statistical methods and results on model performance, validation and impact assessment. Chapter 8 presents a study on the value of a specific method, i.e. principal components analysis, for the reduction of the number of candidate predictors for inclusion in the multivariable modeling. Principal components analysis is a strategy that summarizes multiple predictors into fewer components that can subsequently be used when developing a prediction model, particularly in the case of many candidate predictors relative to the number of events. In chapter 9, we compare the development of a prediction model of an outcome that is inherently continuous both without dichotomisation (using linear regression) and after dichotomisation (using logistic regression), and compared the external validity of the two different models.
Chapter 10 provides concluding remarks and perspectives on future research.

References

(1) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338: b375.

(2) Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

(3) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.

(4) Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338: b606.

(5) Altman DG, Lyman GH (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 52: 289-303.

(6) Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8: 21.

(7) Mallett S, Royston P, Dutton S, Waters R, Altman DG (2010) Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 8: 20.

(8) Hayden JA, Cote P, Bombardier C (2006) Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 144: 427-437.

(9) Perel P, Edwards P, Wentz R, Roberts I (2006) Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 6: 38.

(10) Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ (2011) Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Clin Invest 41: 1004-1009.

(11) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(12) Laupacis A, Sekar N, Stiell IG (1997) Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 277: 488-494.

(13) Toll DB, Janssen KJ, Vergouwe Y, Moons KG (2008) Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 61: 1085-1094.

(14) Wasson JH, Sox HC, Neff RK, Goldman L (1985) Clinical prediction rules. Applications and methodological standards. N Engl J Med 313: 793-799.

(15) Moons KG, Grobbee DE (2002) Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 56: 337-338.

(16) Spiegelhalter DJ (1986) Probabilistic prediction in patient management and clinical trials. Stat Med 5: 421-433.

(17) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(18) Steyerberg, E. W. (2009) Clinical prediction models; a practical approach to development, validation, and updating. New York: Springer.

(19) Van Houwelingen JC, Le Cessie S. (1990) Predictive value of statistical models. Stat Med 9: 1303-1325.

(20) Chatfield C (1995) Model Uncertainty, Data Mining and Statistical Inference. J R Statist Soc A 158: 419-466.

(21) Copas JB (1983) Regression, Prediction and Shrinkage. J R Statist Soc B 45: 311-354.

(22) Rothwell PM (2008) Prognostic models. Pract Neurol 8: 242-253.

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

(24) Concato J, Feinstein AR, Holford TR (1993) The risk of determining risk with multivariable models. Ann Intern Med 118: 201-210.

(25) Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59: 1087-1091.

(26) Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142: 1255-1264.

(27) Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48: 1503-1510.

(28) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379.
(29) Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 56: 441-447.
(30) Gorelick MH (2006) Bias arising from missing data in predictive models. J Clin Epidemiol 59: 1115-1123.
(31) Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. (2006) Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 59: 1092-1101.
(32) van der Heijden GJ, Donders AR, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59: 1102-1109.
(33) Steyerberg EW, Eijkemans MJ, Habbema JD (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 52: 935-942.
(34) Moons KG, Donders AR, Steyerberg EW, Harrell FE (2004) Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 57: 1262-1270.

Part I:
Clinical applications

A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study

Nicolaas P.A. Zuithoff, MSc[1], Yvonne Vergouwe, PhD[1], Michael King, MD, PhD[2] ,Irwin Nazareth, MD, PhD[2], Eelko Hak, PhD[1], Karel G.M. Moons, PhD[1], Mirjam I. Geerlings, PhD[1]

[1]University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, the Netherlands
[2]Department of Mental Health Sciences, Royal Free and University College Medical School, United Kingdom

Published in: Family Practice 2009, june, 26: 241-250

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

Major depressive disorder is a serious health problem. Estimations by the World Health Organization suggest that it will be the second ranking cause of disability by 2020, after cardiovascular disease[1]. Life time prevalence for this mood disorder is estimated up to 25%[2;3]. If not treated, major depressive disorder has a marked impact on quality of life and use of health care services[4-6]. Hence, detection of major depressive disorder is important to improve the patient's prognosis and reduce health care consumption[7;8].

Recent studies showed that a significant proportion of major depressive disorders remain unrecognized by clinicians. Estimates of undetected and therefore untreated depressive disorder are reported over 50%, depending on the studied patient population[5;9-12]. Many instruments are available, designed to screen for major depressive disorder in individual patients[13-16]. However, the majority of these instruments are questionnaires that have to be filled in by the patients themselves, thus requiring an active and relative time consuming strategy to detect primary care patients at high risk of depression.

We aimed to develop a clinical prediction rule to enable primary care physicians to better identify (adult) patients at high risk of major depressive disorder (who may need a more comprehensive diagnostic workup), with minimal involvement of the patients. Hence, we required that the tool can easily be incorporated into the daily routine of the general practitioner. Consequently, the predictors in the clinical prediction rule had to be comprised of patient characteristics or information that will be directly available to the general practitioner.

Methods

Design

This study is part of the international PREDICT study that is set in the Netherlands[17]. In brief, PREDICT is a large prospective cohort study in 6 European countries, the aim of which is to develop a multifactor risk algorithm for the onset of major depression over 12 months in primary care[18]. This study uses the Dutch patients that were included in PREDICT (PREDICT-NL). In the Netherlands, patients were recruited from seven general practices in the city of Utrecht and surrounding areas. Patients aged 18 and over who visited the general practitioner were asked to participate in the study while waiting to see their doctor, irrespective of their reasons for consulting the general practitioner. Patients willing to participate were asked to complete the baseline questionnaire and sign the informed consent form within two weeks. If necessary, a first reminder was sent after two weeks, and a second one after four weeks. Participants who did not respond to the second reminder were considered non-responders. The study was approved by the Medical Ethics committee of the universities of participating countries, and for the Netherlands part by the University Medical Center Utrecht.

The baseline questionnaire was primarily used for the main PREDICT study to collect information on candidate predictors of the patients' prognosis, and included questions about demographics, health, lifestyle, and several psychological measurements. After the informed consent and baseline questionnaire were returned, an appointment was made to conduct the CIDI interview. The CIDI interview was administered separately from the questionnaire. Information regarding health and health care consumption (e.g. consultation rate), however, was extracted from patients medical database, as this was considered a more reliable source.

In total, 3089 patients were asked to take part in the PREDICT-NL study; 75 were excluded because of problems understanding the Dutch language, 5 because of dementia, 2 because of psychosis and 1 because of mental retardation. Of the 3006 eligible patients, 1338 (44.5%) participated in the PREDICT-NL study. Reasons for not participating were mostly lack of time and no interest in the study. Since in the elderly different factors (e.g., cognitive dysfunctioning and functional limitations) may predict the presence or absence of depression as compared to patients at younger age, we included patients aged 18-65 years (n=1046) in the present analysis.

Diagnosis of major depressive disorder

The diagnosis of major depressive disorder at the baseline visit was assessed in all patients according to DSM-IV criteria[19] using the depression section of the Composite International Diagnostic Interview (CIDI)[20]. The CIDI is a structured interview that is administered by trained researchers, it was conducted at the general practice. If the participant was unable to schedule the interview at the general practice, the interview was done by telephone (26% of the interviews) to obtain as complete as possible outcome information. Previous studies showed that telephone interviews are valid for clinical assessment of depression[3;21].

The depression section of the CIDI interview was used to establish whether or not the patient had suffered a major depressive disorder over the past 6 months.

The interviewers were unaware of the values of the diagnostic predictors under study (see below) and the general practitioners were unaware of the diagnosis according to the CIDI interview.

Predictors for the presence of major depressive disorder

We a-priori selected candidate predictors for the presence of major depressive disorder based on the literature and clinical reasoning. These predictors were collected for each patient with the baseline questionnaire and the medical records of the general practitioners. Dutch general practitioners register all contacts, diagnoses, and interventions in an automated database using the International Classification of Primary Care (ICPC)[22;23].

The predictors were divided into two main categories: the first category included easily obtainable predictors that require no sensitive, depression related questions during the consultation. These predictors were mainly obtained from the general practitioner's automated database. The second category included predictors that are known risk factors for and therefore more explicitly related

to, major depressive disorder. The first category of predictors included (1) gender; (2) age; (3) educational level; (4) being single[24;25]; (5) number of presenting complaints at the consult when recruitment took place (inclusion consult)[26;27] (6) assignment of a complaint (ICPC code levels lower than 70) versus a diagnosis (ICPC code levels 70 and higher) at inclusion consult[4] (7) assignment of non-somatic complaint or diagnosis at inclusion consult (ICPC code in chapter A (general), Z (social), or P (psychological) versus any other ICPC chapters, with exclusion of the depression codes P03 and P76)[4]; (8) consultation rate (number of consults in the previous twelve months)[26]; (9) assignment of an ICPC code for depression or depressive complaints (i.e. P03 or P76) in the previous twelve months; (10) prescription of antidepressants in the previous twelve months.

The second category of predictors included the number of life events in the previous six months[28;29] and life time history of depression beyond the previous six months assessed with the two life time questions of the CIDI, i.e. depressed mood and loss of interest for a two week period or longer, ever[28].

Age and consultation rate were analysed as continuous variables. We verified whether these predictors were linearly associated with the outcome, using restricted cubic splines[30]. Educational level was dichotomized into no or primary education only, versus secondary and higher education. The number of complaints presented to the general practitioner at the inclusion consult was categorized into 1, 2 and 3 or more health complaints as only 5.5% of the patients reported 3 or more complaints. The number of life events was categorized into 0, 1, 2, and 3 or more events since only a small number of patients (8%) reported more than 3 life events.

Data analysis

The overall percentage of missing values was 5.9%. Missing data rarely occur at random and a complete case analysis (deletion of all patients with one or more missing values) leads to loss of statistical power and to biased results. We therefore used multiple imputation to address the missing values, including missing values of the outcome[31-33].

Univariable associations between the candidate predictors and the presence of major depressive disorder (yes/no) were estimated with logistic regression analysis. No selection was made based on these estimations, since selection of predictors based on univariable statistics may result in unstable prediction models[30;34].

Selection of predictors was performed in two steps with backward stepwise selection in multivariable logistic regression models. First, the most important predictors of the easily obtainable candidate predictors were selected with age and gender always retained in the model (model 1). Second, the three known risk factors (number of life events in the previous six months, life time depressed mood and life time loss of interest) were added to model 1 to quantify the added diagnostic value (model 2). Backwards selection was based on Akaike's Information Criterium[35], which is similar to a selection based on a p-value of 0.157 if the predictor is modeled with one regression coefficient .

The ability to discriminate between patients with and without a major depressive disorder was studied with the concordance-statistic (c-statistic), i.e. the area under the Receiver Operating Characteristic curve. Calibration, which is the agreement between the observed proportions of major depressive disorder and the predicted risks, was studied with a calibration plot[30;36].

Prediction models derived with multivariable regression analysis are known for overestimated regression coefficients, which results in too extreme predictions when applied in new patients.[30;37]. Therefore, we (internally) validated our models with bootstrapping techniques where in each bootstrap sample the entire modeling process was repeated. This yielded a shrinkage factor for the regression coefficients[30]. The bootstrap procedure was also used to estimate a value of the c-statistic that was corrected for optimism. The corrected c-statistic may be considered as an estimate of discriminative ability that is expected in future patients.

To construct an easy to use clinical prediction rule, the shrunken regression coefficients of the predictors in model 1 and model 2 were transformed into points by multiplying by 10. Coefficients for categorized predictors were then rounded. Coefficients for continuous variables were first multiplied with the variable value and then rounded. The total scores were linked to the risk of major depressive disorder. The analyses were performed with SPSS 14 (SPSS inc., Chicago, Ill, USA) and S-plus 6.2 (Insightful Corp., Seattle, Wa, USA).

Results

The mean age of the 1046 patients was 45 years (SD=13, range 18-65 years), and 673 (64%) were female (Table 1). The majority of patients (n=829, 79%) consulted the general practitioner for one complaint. In 378 patients (36%) the general practitioner did not assign a diagnosis (ICPC-coding below level 70). The median consultation rate in the past 12 months was 8 (interquartile range: 5-15). One hundred and eleven (11%) patients had a non-somatic diagnosis or complaint (ICPC chapters P, Z, or A). Major depressive disorder according to DSM-IV criteria was diagnosed in 157 patients (prevalence or a-priori risk of 15%).

Table 1 Characteristics of 1046 primary care patients. Values are N (%) unless stated otherwise

Characteristics	
Candidate predictors	
Female gender	673 (64)
Age, years[1]	44.7 (12.8)
Educational level, none/primary only	211 (20)
Being single	237 (23)
Number of presented complaints	
1	829 (79)
2	162 (16)
3 or more	55 (5)
General practitioner did not assign a diagnosis at inclusion consult*	378 (36)
Non-somatic diagnosis/complaint at inclusion consult**	111 (11)
Consultation rate (number of consults in past 12 months)[2]	8 (5-15)
Received depression code in past 12 months***	58 (6)
Prescription of antidepressants in past 12 months	90 (9)
Number of life events in past six months	
0	408 (39)
1	285 (27)
2	188 (18)
3 or more	165 (16)
Any depressed feelings, life time	514 (49)
Any loss of interest, life time	421 (40)
Outcome	
Major depressive disorder	157 (15)

[1]mean (SD); [2] median (interquartile range).
* ICPC-coding below level 70
** ICPC chapters general (A), social (Z) or psychological (P) excluding codes P03 and P76.
*** ICPC codes P03 or P76.

Female gender was univariably associated with a higher risk of having major depressive disorder (Table 2). Other variables that were univariably associated with a high risk of major depressive disorder were younger age, low educational level, being single, more than one presenting complaints, non-somatic diagnosis or complaint, higher consultation rate, depression code (P03 or P76) in past 12 months, prescription of antidepressants in past 12 months, one or more life events in the preceding six months and life time history of depression (Table 2).

Multivariable regression analysis showed that all easily obtainable predictors - except complaint versus diagnosis - remained in the model (model 1, Table 3). The *c*-statistic of the model was 0.71 (95% CI: 0.67-0.76). Model extension with the three additional predictors - number of life events in the previous six months and the two lifetime questions on history of depression increased the discriminative ability of the model to a *c*-statistic of 0.80 (95% CI: 0.76-0.83) (model 2, Table 3). The effect of gender was retained in the model, even though the odds ratio was reduced to nearly 1 (model 2, Table 3).

Table 2 Univariable associations between candidate predictors and major depressive disorder. Values are N (%) unless stated otherwise

Predictor	Major depressive disorder		
	Yes	No	Odds Ratio (95% CI)
Female gender	115 (73)	558 (63)	1.62 (1.11-2.37)
Age, years[1]	43.1 (12)	45.0 (13)	0.99 (0.98-1.00)
Educational level, none/primary only	44 (28)	167 (19)	1.54 (0.97-2.46)
Being single	55 (35)	182 (21)	2.10 (1.45-3.03)
Number of presented complaints			
1	107 (68)	722 (81)	╫
2	34 (22)	128 (14)	1.65 (1.07-2.56)
3 or more	16 (10)	39 (4)	2.75 (1.48-5.10)
General practitioner did not assign a diagnosis at inclusion consult*	50 (32)	328 (37)	0.89 (0.59-1.34)
Non-somatic diagnosis/complaint at inclusion consult**	29 (19)	82 (9)	2.21 (1.39-3.51)
Consultation rate (number of consults in past 12 months)[2]	11 (7-20)	8 (5-14)	1.06 (1.04-1.09)
Received depression codes in past 12 months***	33 (21)	25 (3)	8.94 (5.13-15.6)
Prescription of antidepressants in past 12 months	41 (26)	49 (6)	6.03 (3.81-9.54)
Number of life events in past six months			
0	36 (23)	372 (42)	╫
1	33 (21)	252 (28)	1.36 (0.83-2.25)
2	27 (17)	161 (18)	1.71 (1.00-2.93)
3 or more	61 (39)	104 (12)	6.09 (3.82-9.70)
Any depressed feelings, life time	120 (76)	394 (44)	4.07 (2.75-6.02)
Any loss of interest, life time	108 (69)	313 (35)	4.04 (2.80-5.83)

[1]mean (SD); [2]median (interquartile range); * ICPC-coding below level 70; ** ICPC chapters general (A), social (Z) or psychological (P) excluding codes P03 and P76; *** ICPC codes P03 or P76; ╫ reference category

Table 3 Multivariable logistic regression models for the diagnosis of major depressive disorder.

Predictor	Model 1 Odds Ratio (95%CI)	Beta[1] (P value)	Model 2 Odds Ratio (95%CI)	Beta[1] (P value)
Female gender	1.20 (0.79-1.82)	0.18 (0.40)	1.01 (0.64-1.57)	0.01 (0.98)
Age, per year increase	0.99 (0.97-1.03)	-0.01 (0.08)	0.99 (0.97-1.00)	-0.01 (0.07)
Educational level, none/primary only	1.50 (0.97-2.33)	0.41 (0.07)	1.42 (0.90-2.26)	0.35 (0.14)
Being single	1.52 (1.00-2.29)	0.42 (0.05)	1.27 (0.82-1.96)	0.23 (0.29)
Number of presented complaints				
1	╪	╪	╪	╪
2	1.46 (0.91-2.37)	0.38 (0.12)	1.37 (0.83-2.26)	0.31 (0.23)
3 or more	2.03 (1.02-4.03)	0.71 (0.04)	2.09 (1.01-4.33)	0.74 (0.05)
Non-somatic diagnosis/complaint at inclusion consult*	1.63 (0.95-2.79)	0.49 (0.07)	1.62 (0.89-2.95)	0.48 (0.11)
Consultation rate (number of consults in past 12 months)	1.03 (1.00-1.06)	0.03 (0.03)	1.02 (0.99-1.05)	0.02 (0.27)
Received depression code in past 12 months**	3.52 (1.77-6.99)	1.26 (0.00)	3.26 (1.59-6.70)	1.18 (0.00)
Prescription of antidepressants in past 12 months	2.33 (1.29-4.20)	0.85 (0.00)	2.00 (1.07-3.76)	0.70 (0.03)
Number of life events	-	-		
0			╪	╪
1			1.21 (0.70-2.08)	0.19 (0.49)
2			1.41 (0.79-2.53)	0.34 (0.25)
3 or more			3.72 (2.20-6.29)	1.31 (0.00)
Any depressed feelings, life time	-	-	1.71 (0.99-2.97)	0.54 (0.06)
Any loss of interest, life time	-	-	1.70 (1.01-2.86)	0.53 (0.04)

[1] Beta: logistic regression coefficient, which was shrunk to improve predictions for future patients
* ICPC chapters general (A), social (Z) or psychological (P), excluding codes P03 and P76
** ICPC codes P03 or P76; ╪ reference category; - not selected in the model

Figures 1 and 2 show the agreement between the predicted risks estimated with model 1 and 2, respectively, and the observed proportions of major depressive disorder. Predicted risks around 25% of model 1 were underestimated where predicted risks of 55% and higher were overestimated. Model 2 clearly showed better calibration, with some discrepancy between predicted and observed risk in the high range of predicted risks (>50%). This is largely due to the low number of patients in these groups. Figures 3 and 4 show the score chart derived from model 1 and 2 respectively (Table 3) that can be used as a clinical prediction rules. The regression coefficient for gender in model 2 was close to zero (0.01). Therefore, gender was not included in the score chart. The lower part of Figures 3 and 4 show predicted risks and observed proportions for ranges of total scores. As with the calibration plots, some discrepancies between predicted risks and observed proportions were observed, especially in higher predicted risk categories, where the number of patients is low. An example of the use of the clinical prediction rules is given in the legend.

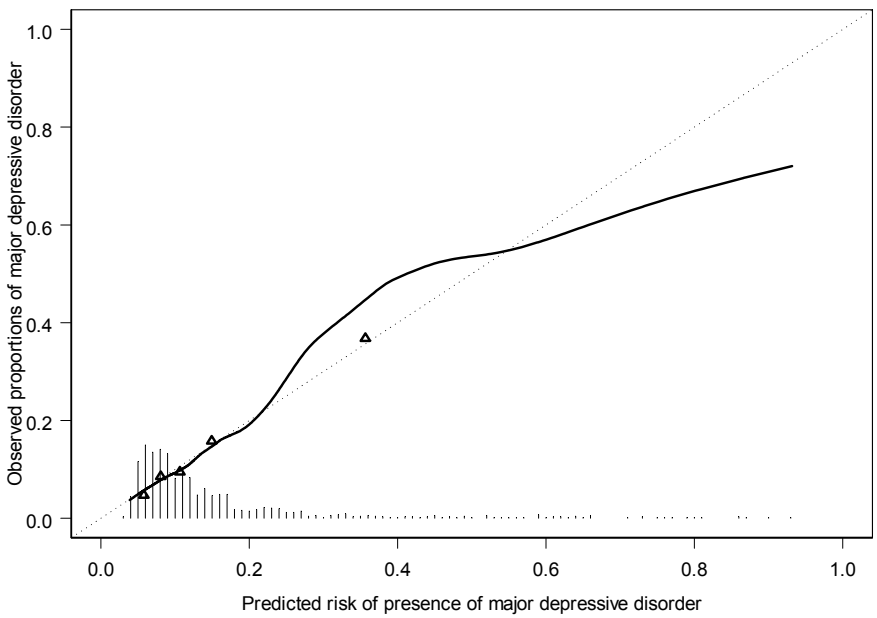


Figure 1.
Agreement between the predicted risks of major depressive disorder according to model 1 and the observed proportions. The solid line indicates the agreement between predicted risks of major depressive disorder and observed proportions. The dotted line indicates ideal calibration. The triangles indicate the observed proportions of major depressive disorder in patients with similar predicted risks grouped in quintiles. The vertical lines just above the horizontal axis show the distribution of the predicted risks.

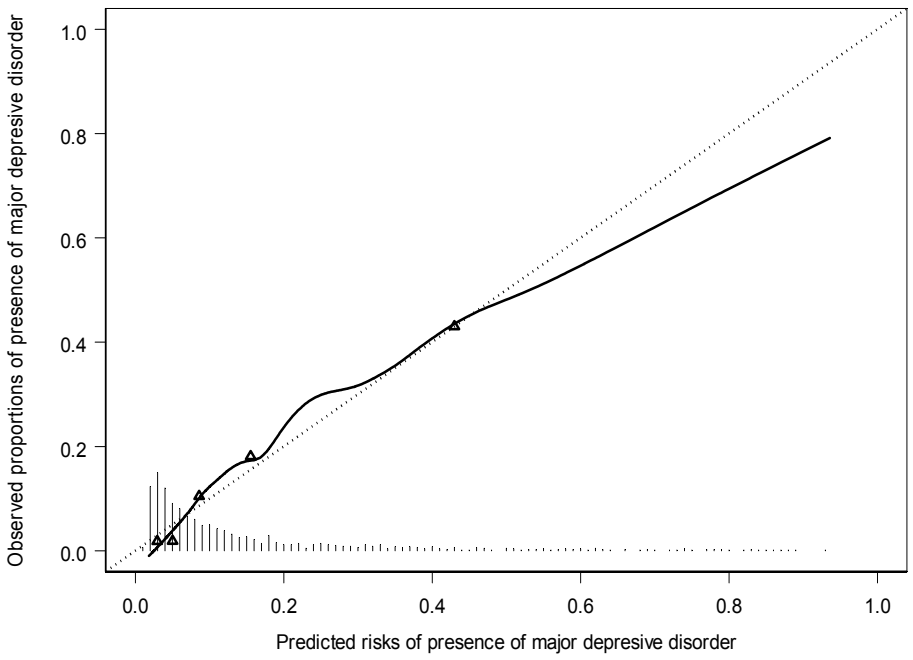


Figure 2.
Agreement between the predicted risks of major depressive disorder according to model 2 and the observed proportions. The solid line indicates the agreement between predicted risks of major depressive disorder and observed proportions. The dotted line indicates ideal calibration. The triangles indicate the observed proportions of major depressive disorder in patients with similar predicted risks grouped in quintiles. The vertical lines just above the horizontal axis show the distribution of the predicted risks.

Points	0	1	2	3	4	5	6	7	8	13
Female gender	M		F							
Age, years[1]	58	50	43	35	27	20	18			
Educational level, none/primary only	No				Yes					
Being single	No				Yes					
Number of presenting complaints	1				2			3+		
Non-somatic diagnosis or complaint at inclusion consultation*	No					Yes				
Consultation rate, number of consultations in past 12 months[1]	0	2	5	8	11	15	18	21	24	
Received depression code in previous 12 months**	No									Yes
Prescription of antidepressants in previous 12 months	No								Yes	

[1] Choose the highest number of points when the patient's value falls between the given values
* ICPC chapters general (A), social (Z) or psychological (P), depression codes P03 and P76 not included
** ICPC codes P03 or P76

Total score	Predicted risk, %	Observed proportion, % (n/N) **
0-5	6	6 (15/250)
6-10	9	10 (38/388)
11-15	14	14 (31/225)
16-20	21	24 (22/90)
21-25	30	51 (18/35)
26+	52	57 (33/58)
Total	15	15 (157/1046)

*percentage of the "Total" column
**number of patients with major depressive disorder/total number of patients in this risk category

Figure 3.
Score chart based on model 1 to calculate the predicted risk of major depressive disorder for an individual primary care patient. The upper part shows the points corresponding to each predictor value. For the continous predictors not all values are given. The correct number of points for age can be found by rounding downwards to a value in the chart. The correct number of points for consultation rate can be found by rounding upwards to a value in the chart. The points are summed up into a score. The corresponding risk for major depressive disorder can be found for ranges of scores in the lower part of Figure 2 in the column Predicted risk. For comparison, the observed percentage of patients with major depressive disorder is shown in the column Observed proportions.

Points	0	1	2	3	4	5	6	7	8	12	13
Age, years[1]	58	50	43	35	27	20	18				
Educational level, none/primary only	No				Yes						
Being single	No		Yes								
Number of presenting complaints	1			2				3+			
Non-somatic diagnosis or complaint at inclusion consultation*	No					Yes					
Consultation rate, number of consults in past 12 months[1]	0	2	5	8	11	15	18	21	24		
Received depression code in previous 12 months**	No									Yes	
Prescription of antidepressants in previous 12 months	No							Yes			
Number of life events	0		1	2							3+
Any depressed feelings, life time	No					Yes					
Any loss of interest, life time	No					Yes					

[1] Choose the highest number of points when the patient's value falls between the given values
* ICPC chapters general (A), social (Z) or psychological (P), depression codes P03 and P76 not included
** ICPC codes P03 or P76

Total score	Predicted risk, %	Observed proportion, % (n/N)
0-5	3	1 (2/139)
6-10	5	1 (3/228)
11-15	8	9 (19/215)
16-20	13	15 (26/176)
21-25	20	21 (22/104)
26-30	29	43 (35/89)
30+	52	49 (50/103)
Total	15	15 (157/1046)

*percentage of the "Total" column

Figure 4.
Score chart based on model 2 to calculate the predicted risk of major depressive disorder for an individual primary care patient. The upper part shows the points corresponding to each predictor value. For the continous predictors not all values are given. The correct number of points for age can be found by rounding downwards to a value in the chart. The correct number of points for consultation rate can be found by rounding upwards to a value in the chart. The points are summed up into a score. The corresponding risk for major depressive disorder can be found for ranges of scores in the lower part of Figure 3 in the column Predicted risk. For comparison, the observed percentage of patients with major depressive disorder is shown in the column Observed proportions.
To illustrate the use of the score chart: A patient, aged 27 (4 points), with a high education (points) and single (2 points) consults the general practitioner. The patient presents 3 separate health complaints (7 points), all somatic (0 points). This consultation is the 3rd in the past twelve months (2 points). The medical database shows that the patient did not receive a depression code in the prior consultations (0 points) and no anti-depressive medication was prescribed in the previous 12 months (0 points). The general practitioner further inquires about recent life events (3 life events, 13 points) and asks whether the patient has ever had depressed feelings for more than 14 days (yes, 5 points) or loss of interest (no, 0 points). The score is 33, which relates to a predicted risk of major depressive disorder of 52% (Figure 4, lower part).

Discussion

We developed a two-step clinical prediction rule to predict the likelihood of presence of major depressive disorder in primary care patients. The first step was to develop a model with easily obtainable predictors (model 1, Table 3) that could be used by the general practitioner without asking for specific, sensitive questions with reference to major depressive disorder. This was done because general practitioners may be reluctant to ask specific depression related questions, particularly if patients present with somatic complaints. In addition, patients may be reluctant to answer such questions. This first model was then extended by inclusion of three known risk factors that are more explicitly related to major depressive disorder (model 2). This extension clearly improved the discriminative ability and calibration of the model. The use of both models is facilitated by a simple to use score chart (figure 3 and 4). Although the model requires use of more than 10 parameters, the majority of predictors can simply be derived from the (often electronic) medical files in primary care practices, without having to question the patient. The applicability of the two-step clinical prediction rule can obviously be improved by incorporating it (notably the first model) as an automatic tool (calculator) in the electronic patient medical record. During the consultation the program could 'warn' general practitioners if a patient is at an elevated risk of major depressive disorder.

Since adding the three extra predictors in model 2 substantially improved the discrimination, we advocate a two-step approach to select patients in primary care practice who can benefit from diagnostic workup. Model 1 can be used as a first selection tool to distinguish patients with lower predicted risk from patients with higher predicted risk of major depressive disorder. The use of model 1 can be facilitated by a computer program (using data from the most recent consultation for the "inclusion consult") to calculate the predicted risk and alert the general practitioner to patients at relatively high risk during the consult. The doctor might only ask questions related to major depressive disorder that are included in model 2 when the predicted risk of model 1 is sufficiently high. For example, a female patient (2 points), aged 25 (5 points), with primary education only (4 points) and single (4 points) consults the general practitioner. The patient presents with 3 health complaints (7 points), all somatic (0 points). This consultation is the second in the past twelve months (1 point) and the visits did not result in a ICPC code for depression (0 points) or in prescriptions for anti-depressants (0 points). As a consequence, this patient receives a total score of 23 that relates to a predicted risk of 30% (see appendix). For this patient, it seems reasonable to ask the additional questions of model 2.

We need to define threshold values for 'high risk of major depressive disorder', in order to apply the two models. Ideally, such threshold values are assessed in formal decision analyses that weigh benefits and harms of further diagnostic-work up. Since these analyses are currently lacking, we propose a low threshold of 11 or higher for model 1. Using this threshold 66% of the patients with major depressive disorder will be asked the additional questions of model 2. For model 2, we

would recommend a threshold of 21 or higher as an indicator for a further diagnostic work-up. This would result in a diagnostic work-up for most of the patients with major depressive disorder (n=107; 68%). We report the risk for several total score categories of both models to enable general practitioners to choose their desired threshold of risk of major depressive disorder.

Strong predictors (variables with high odds ratios) in model 1 were the assignment of an ICPC code for depression or depressive complaint in the past twelve months, and prescription of antidepressants in the past twelve months. The number of presented health complaints and assignment of a non-somatic diagnosis or complaint at the inclusion consult were also strong predictors. These results are consistent with other studies[25;26;28].

It may seem odd to include the predictors 'depression or depressive complaint in the last 12 months' and 'prescription of antidepressants in the last 12 months' for the detection of major depressive disorder. However, a diagnosis of major depressive disorder by general practitioners is usually not assessed with reference tests such as the CIDI, and therefore not the same as the diagnosis of major depressive disorder assessed in our study. Further, antidepressants may be used for other conditions (i.e. anxiety, obsessive compulsive disorder, dysthymia) than major depressive disorder.

Three or more life events increased the risk for major depressive disorder dramatically. Life events are a known risk factor for major depressive disorder[11;24;25]. The clear distinction in risk between one or two versus three or more life events was unexpected, though it has been reported that a high number of life events is associated with persistence of major depressive disorder[38]. Gender was predictive in model 1, but had no predictive value anymore in model 2. This is due to the association between gender and the extra three predictors that were included in model 2, which in fact took over the predictive ability of gender. Women had experienced life events more often and responded positively more often on the CIDI lifetime questions.

The question may arise whether our models were similar if patients who were already diagnosed by their general practitioner for depression were excluded from the analysis. An additional analysis excluding patients with ICPC depression codes P03 or P76, showed similar results (analysis not shown). The selected predictors were the same - except for the ICPC depression codes that were excluded in this analysis - and yielded similar regression coefficients. We specifically chose to include patients with recognized or treated depression, since the degree of recognition might vary across general practitioners.

To our knowledge, this is the first attempt to develop and internally validate a simple rule for general practitioners to determine the risk of major depressive disorder in individual patients, with a minimal need for depression related questions. A strength of this study is the inclusion of consecutive primary care patients, irrespective of their presented symptoms or signs. No bias was introduced due to selection, as the reference test was applied in all patients.

Our study has some limitations. First, the non-response rate for our study was relatively high. Compared to responders, however, we found very minor differences in distributions of gender and age. Since other predictors could not be collected in the non-response groups, we can only assume that both groups were largely similar with respect to the other characteristics. Furthermore, the prevalence of major depressive disorder in this study was relatively high. This may suggest that patients with mood problems were more willing to participate and may have resulted in an overestimation of the predicted risk of major depressive disorder. The low response rate and high prevalence could indeed be indications of some bias in the inclusion of patients. Consequently, we stress that external validation of our prediction rule, as is always the case with developed rules, is needed to study the generalisability of our models.

Second, some predictors in model 1 may currently not be available on GP records. For example in the UK the information on single status and education is not available. General practitioners should first have this information, before the model can be automatically applied. Third, the clinical prediction rule was explicitly developed on data of adult patients aged 18 to 65. Older patients were excluded, the prediction rule may not be valid for these patients.

In conclusion, we have developed a clinical prediction rule to detect major depressive disorder in adult primary care patients. The two score charts were internally validated with a bootstrapping technique. External validation, including prospective testing, is required before the models can be applied in other general practices. If proved to be generalisable, the prediction rules can be used to detect primary care patients for a work-up to diagnose major depressive disorder. Early detection of major depressive disorder, provided it is followed by adequate treatment and follow-up, may improve the long-term prognoses for these patients.

References

(1) Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. Lancet 1997 May 17;349(9063):1436-42.

(2) Kruijshaar ME, Barendregt J, Vos T, De Graaf R, Spijker J, Andrews G. Lifetime prevalence estimates of major depression: an indirect estimation method and a quantification of recall bias. Eur J Epidemiol 2005;20(1):103-11.

(3) Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003 Jun 18;289(23):3095-105.

(4) Goldman LS, Nielsen NH, Champion HC. Awareness, diagnosis, and treatment of depression. J Gen Intern Med 1999 Sep;14(9):569-80.

(5) McQuaid JR, Stein MB, Laffaye C, McCahill ME. Depression in a primary care clinic: the prevalence and impact of an unrecognized disorder. J Affect Disord 1999 Sep;55(1):1-10.

(6) Rost K, Zhang M, Fortney J, Smith J, Coyne J, Smith GR, Jr. Persistently poor outcomes of undetected major depression in primary care. Gen Hosp Psychiatry 1998 Jan;20(1):12-20.

(7) Pignone MP, Gaynes BN, Rushton JL, Burchell CM, Orleans CT, Mulrow CD, et al. Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002 May 21;136(10):765-76.

(8) Bijl D, van Marwijk HW, de Haan M, van Tilburg W, Beekman AJ. Effectiveness of disease management programmes for recognition, diagnosis and treatment of depression in primary care. Eur J Gen Pract 2004 Mar;10(1):6-12.

(9) Barkow K, Maier W, Ustun TB, Gansicke M, Wittchen HU, Heun R. Risk factors for depression at 12-month follow-up in adult primary health care patients with major depression: an international prospective study. J Affect Disord 2003 Sep;76(1-3):157-69.

(10) Nuyen J, Volkers AC, Verhaak PF, Schellevis FG, Groenewegen PP, Van den Bos GA. Accuracy of diagnosing depression in primary care: the impact of chronic somatic and psychiatric co-morbidity. Psychol Med 2005 Aug;35(8):1185-95.

(11) Olsen LR, Mortensen EL, Bech P. Prevalence of major depression and stress indicators in the Danish general population. Acta Psychiatr Scand 2004 Feb;109(2):96-103.

(12) Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A. Recognition of depression by non-psychiatric physicians--a systematic literature review and meta-analysis. J Gen Intern Med 2008 Jan;23(1):25-36.

(13) Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-13.

(14) Williams JW, Jr., Pignone M, Ramirez G, Perez SC. Identifying depression in primary care: a literature synthesis of case-finding instruments. Gen Hosp Psychiatry 2002 Jul;24(4):225-37.

(15) Williams JW, Jr., Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? JAMA 2002 Mar 6;287(9):1160-70.

(16) Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-70.

(17) King M, Weich S, Torres F, Svab I, Maaroos H, Neeleman J, et al. Prediction of depression in European general practice attendees: the PREDICT study. BMC Public Health 2006 Jan 12;6(1):6.

(18) King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, et al. Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study. Arch Gen Psychiatry 2008 Dec;65(12):1368-76.

(19) American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4th Text Revision ed. Washington, D.C.: American Psychiatric Association; 2000.

(20) Ter Smitten RH, Smeets RMW, Van den Brink W. Composite International Diagnostic Interview - computerized version 2.1: Dutch translation and adaptation. WHO-CIDI Training en Referentie centrum, Psychiatrisch centrum AMC: Amsterdam; 2007.

(21) Sobin C, Weissman MM, Goldstein RB, Adams P, Wickramaratne P, Warner V, et al. Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. Psychiatric Genetics 1993;3(4):227-34.

(22) Classification Committee of WONCA. International Classification of Primary Care. Oxford: Oxford University Press; 1998.

R1 (23) Okkes I, Jamoulle M, Lamberts H, Bentzen N. ICPC-2-E: the electronic version of ICPC-2. Differences from the printed version and the consequences. Fam Pract 2000 Apr;17(2):101-7.
R2 (24) De Graaf R, Bijl RV, Ravelli A, Smit F, Vollebergh WA. Predictors of first incidence of DSM-III-R psychiatric disorders in the general population: findings from the Netherlands Mental Health Survey and Incidence Study. Acta Psychiatr Scand 2002 Oct;106(4):303-13.
R3 (25) Salokangas RK, Poutanen O. Risk factors for depression in primary care. Findings of the TADEP project. J Affect Disord 1998 Mar;48(2-3):171-80.
R4 (26) Katon W, Berg AO, Robins AJ, Risse S. Depression--medical utilization and somatization. West J Med 1986 May;144(5):564-8.
R5 (27) Ormel J, Bartel M, Nolen WA. [Undertreatment of depression; causes and recommendations]. Ned Tijdschr Geneeskd 2003 May 24;147(21):1005-9.
R6 (28) Barkow K, Maier W, Ustun TB, Gansicke M, Wittchen HU, Heun R. Risk factors for new depressive episodes in primary health care: an international prospective 12-month follow-up study. Psychol Med 2002 May;32(4):595-607.
R7 (29) Brugha TS, Cragg D. The List of Threatening Experiences: the reliability and validity of a brief life events questionnaire. Acta Psychiatr Scand 1990 Jul;82(1):77-81.
R8 (30) Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.
R9 (31) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.
R10 (32) Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.
R11 (33) van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006 Oct;59(10):1102-9.
R12 (34) Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996 Aug;49(8):907-16.
R13 (35) Atkinson A. A note on the generalized information criterion for choice of a model. Biometrika 1980;67:413-8.
R14 (36) Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991 Aug;10(8):1213-26.
R15 (37) Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986 Sep;5(5):421-33.
R16 (38) Spijker J, De Graaf R, Bijl RV, Beekman AT, Ormel J, Nolen WA. Determinants of persistence of major depressive episodes in the general population. Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). J Affect Disord 2004 Sep;81(3):231-40.
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

3

The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study

Nicolaas P.A. Zuithoff, MSc[1], Yvonne Vergouwe, PhD[1], Michael King, MD, PhD[2], Irwin Nazareth, MD, PhD[2], Manja J. van Wezep, Msc[1,3], Karel GM Moons, PhD[1], Mirjam I. Geerlings, PhD[1]

[1]University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, the Netherlands
[2]Department of Mental Health Sciences, Royal Free and University College Medical School, United Kingdom
[3]Netherlands Institute of Mental Health and Addiction, Trimbos Institute, Utrecht

Published in: BMC Family Practice 2010 11:98.

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Background

Assessment of major depressive disorder with (semi-)structured interviews such as the CIDI or the SCID[1,2], can be time consuming in the primary care setting. There is a need for using brief instruments such as the Patient Health Questionnaire-9 (PHQ-9) and the PHQ-2[3], to ascertain the diagnosis of major depressive disorder.

The PHQ-9 is derived from PRIME-MD[4] which was originally developed to detect five common mental disorders in primary care: depression, anxiety, alcohol abuse, somatoform disorder, and eating disorder. It is a self-report questionnaire that assesses the levels of depression on the nine key symptoms (each rated from 0-3) in the past two weeks. The scores on the questionnaire range from 0 to 27: a score of 10 or higher is indicative of moderate or severe depression and is used to consider major depressive disorder present[3,5-8]. The score can also be used as a measure of depression severity[3,9]. A categorical algorithm has also been developed to determine major depressive disorder with the PHQ-9[3,9]. The PHQ-2 includes the first two items of the PHQ-9, 'any depressed feelings' and 'any loss of interest'[10] and ranges from 0 to 6. In order to detect major depressive disorder with the PHQ-2, a threshold of 3 is recommended.

Several studies validated the performance of both questionnaires in a variety of patient populations, most of them showing good accuracy[3,5-8,11-16]. However, the PHQ-9 has not yet been validated in primary care in the Netherlands. Furthermore, very few studies validated the accuracy of the PHQ-2[10,12,16,17].

We validated both the PHQ-9 and the PHQ-2 in a large Dutch primary care patient cohort addressing three questions: (1) Is the PHQ-9 a reliable and valid measurement of major depressive disorder in primary care? Reliability refers to internal consistency as well as test-retest reliability. Validity refers to construct validity, i.e. is the PHQ-9 an adequate measurement of depression severity; 2) Does the threshold score of 10 and the categorical algorithm for the PHQ-9 yield accurate classification in primary care?; (3) What is the accuracy of the PHQ-2 for major depressive disorder in primary care?

Methods

Patients and design

We used patient data of the PREDICT-NL study, which is the Dutch part of the PredictD study. The design and primary results of the PredictD study have been published previously[18,19]. In brief, PredictD is a large prospective cohort study that started in 2003 from which a multifactor risk algorithm was developed for the onset of major depression over 12 months in primary care in 6 European countries and Chile[18]. Consecutive general practice patients were asked to participate, irrespective of their reasons for consulting the general practitioner. The study was approved by the Medical Ethics committee of the universities of participating countries.

In the Netherlands, patients were recruited from seven general practices in the city of Utrecht and surrounding areas. On random days, research assistants visited the general practices to recruit patients. Patients aged 18 years or older who visited the general practice were asked to participate while waiting to see the general practitioner. Patients interested in participating were given oral and written information about the study aims and procedure. If patients were willing to participate, they received the study information sheet, an informed consent form, and the questionnaires. The patient was asked to take the material home, read the study information and ask for additional information if necessary. After having signed the informed consent form they filled out the questionnaire and returned the signed informed consent form and questionnaire by regular mail. Nonresponders were sent a reminder after two weeks and again after four weeks.
To assess the test-retest reliability of the PHQ-9, thirty-two consecutively included study participants in one general practice were asked to fill out the PHQ questionnaire for a second time after 14 days.

Diagnosis of major depressive disorder (reference standard)

The diagnosis of major depressive disorder was assessed in all patients according to DSM-IV criteria[20] by trained researchers using the depression section of the Composite International Diagnostic Interview (CIDI)[21]. When informed consent and the questionnaire were received, the researchers phoned the participant and asked the two core questions of the depression section of the CIDI interview[21], i.e. did you have a depressed mood or a loss of interest for a 2-week-period or longer in the past six months. If the participant responded negative to both questions a diagnosis of major depressive disorder was ruled out[20,21]. If the participant responded positive on one or both questions, an appointment was made in the general practice to conduct the full CIDI depression interview to establish the presence of major depressive disorder. If the participant was unable to schedule the interview at the general practice, the interview was done by telephone (26% of the interviews). The electronic processing of the questionnaires was done completely separate from the CIDI interview, thus effectively blinding the researchers from the PHQ-9 answers.

Patient health questionnaire

Each of the nine questions of the PHQ-9 was evaluated on a 4-point rating scale, ranging from 0 (not at all) to 3 (nearly every day), summing up to a total PHQ-9 score per patient. Major depressive disorder was considered present if the score was >= 10[3,5-8]. For the categorical algorithm, the answers on the questions were dichotomized: 0 (not at all) and 1 (several days) are coded as 0 (symptom absent) and the answers 2 (more than half the days) and 3 (nearly every day) are coded as 1 (symptom present). The diagnosis of major depressive disorder is made when at least five symptoms are present, and at least one is 'depressed feelings' or 'loss of interest'[3,20].
For the PREDICT-NL study, the Dutch version of the PHQ-9 was developed using several steps of translating and back-translating by researchers and professional translators, one of whom was a

native English speaker. The PHQ-2 is a reduced version of the PHQ-9: only the core symptoms of major depressive disorder ('depressed feelings' and 'loss of interest'), the first two items, are measured as described above, summing up to a total that ranges from 0 to 6.

Functional status, sick days, and number of consultations

We also assessed other parameters to evaluate the validity of the PHQ-9. These were:

1) Functional status using the Medical Outcome Study Short Form General Health Questionnaire-12 (SF-12)[22]. This instrument is divided into scales for mental and physical health, where higher scores indicate better functioning.
2) Information on the number of days in the past 4 weeks that patients were unable to perform usual activities due to health problems (number of sick days).
3) The number of general practice consultations in the past 12 months was counted using the electronic database of the general practitioners. This was assessed as a measure of health service utilisation.

Data analysis

We estimated the internal consistency, the degree to which the answers on the individual questions of the PHQ-9 are the same, of the PHQ-9 using intraclass correlations and the test-retest correlation were estimated using Pearson correlations. To assess the validity of the PHQ-9 as a measurement of depression severity, scores were divided in categories of increasing severity: 0-4, 5-9, 10-14, 15-19 and 20 and higher, as used in other studies[3]. Medians and interquartile ranges of the functional status (SF-12), sick days and the number of consultations in the previous 12 months were estimated across these categories. Differences between categories were tested with Kruskal-Wallis analysis of variance. Differences in PHQ-9 and PHQ-2 scores between patients with and without major depressive disorder were tested with the Mann-Whitney U test. P-values of 0.05 and lower were considered significant.

We then estimated the concordance-statistic (c-statistic or area under the Receiver Operating Characteristic curve) for the PHQ-9. The sensitivity, specificity, and positive and negative predictive value were estimated for several thresholds of the PHQ-9 overall score and for the categorical algorithm of the PHQ-9. Finally , the c-statistic was constructed for the PHQ-2 and sensitivity, specificity, positive and negative predictive value were calculated for all possible thresholds of the PHQ-2.

The overall percentage of missing values was 9%. As missing data rarely occur at random, it is widely acknowledged that simple deletion of patients with one or more missing values (i.e. complete case analysis) leads to biased results[23-26]. We therefore used single imputation to address missing values. The imputation and analysis was done in SPSS version 15 (SPSS inc. Chicago, Ill).

Results

In total, 3089 patients were asked to take part in the PREDICT-NL study, 83 of whom did not meet inclusion criteria, mainly (n=75) because they had problems understanding the Dutch language. An additional 8 patients were excluded because the general practitioner confirmed that they had dementia (n=5), psychosis (n=2), or mental retardation (n=1). Of the 3006 eligible patients, 1338 (44.5%) gave written informed consent and participated in the study (Figure 1). Reasons for not participating were mostly lack of time and no interest in the study.

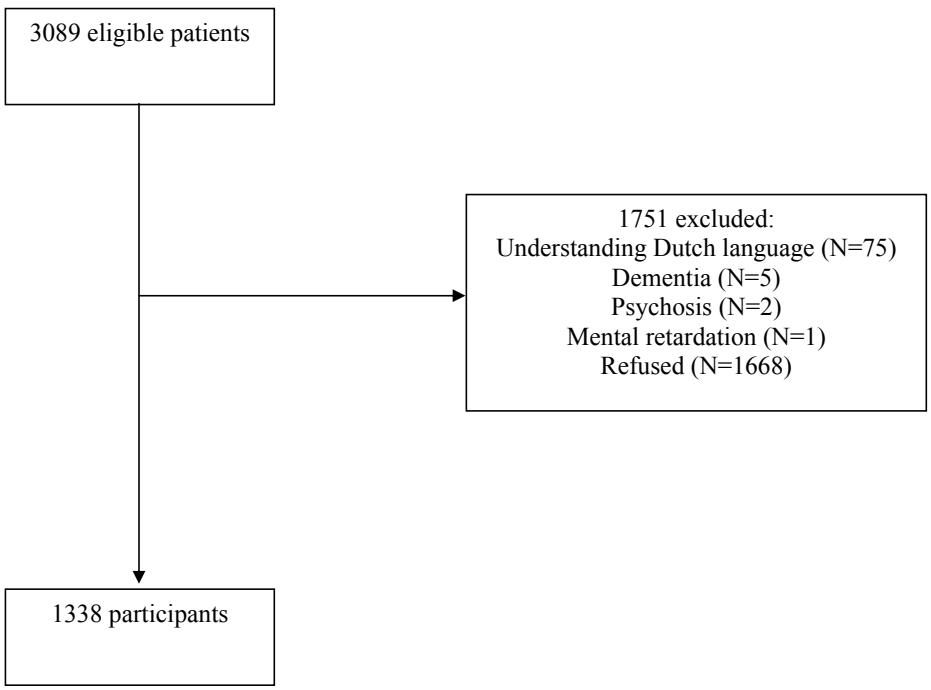


Figure 1.
Flow chart of the inclusion of patients.

The mean age of the study population was 51 years (SD=16.7), and the majority (63%) was female (Table 1). Thirty five patients (2.6%) consulted the general practitioner for mood related health problems. DSM IV Major depressive disorder according to the CIDI was diagnosed in 176 (13%) patients and was more prevalent in women and younger patients (Table 1). Patients with major depressive disorder had significantly higher scores on the PHQ-9 (p<.000) and PHQ-2 (p<.000) compared with patients without depression.

The association between the test and retest scores was excellent, with a correlation of 0.94. Thirty-one of the 32 patients approached agreed to fill in the PHQ-9 for a second time. The internal consistency of the PHQ-9 was very good with an intraclass correlation of 0.88.

Table 1 Distribution of patient characteristics according to diagnostic status for major depressive disorder

	Major depressive disorder		Total
	No (N=1176)	Yes (N=176)	
Male gender, n (%)	448 (39)	50(28)	498 (37)
Age (years), mean(SD)	52 (17)	46(14)	51 (17)
PHQ-9, median(IQR[1])	2 (1-5)	9 (6-14)	3 (1-6)
PHQ-2, median(IQR)	0 (0-1)	2 (2-4)	0 (0-2)
Physical functioning, SF-12, median(IQR)	50 (40-54)	48 (40-57)	49 (40-54)
Mental functioning, SF-12, median(IQR)	52 (45-56)	30 (25-38)	50 (41-56)
Sick days (n), median(IQR)	0 (0-5)	3 (0-10)	0 (0-5)
Consultations in previous 12 months (n), median(IQR)	9 (5-16)	12 (7-20)	10 (5-16)

[1] IQR - interquartile ranges

In table 2, the medians of the SF12, the number of sick days and the number of consultations are shown for patients in different PHQ-9 categories. A statistically significant difference in quality of life was observed for patients with different levels of depressive symptoms, with patients with higher levels of depression reporting a lower quality of life. This difference was more pronounced on the mental functioning than the physical functional scale. A statistical significant difference was also observed on the reported number of sick days in the past 4 weeks and number of consultations in the past twelve months (all p-values < 0.001).

Table 2 Association between PHQ-9 depression score and SF-12 health related quality of life scores, sick days and number of consultations in the past 12 months.

	Level of depression severity (PHQ-9 score)					
	Minimal (0-4)	Mild (5-9)	Moderate (10-14)	Moderately severe (15-19)	Severe (20-27)	p-value[2]
Physical functioning, Median (IQR[1])	50 (42-54)	47 (38-54)	47 (38-57)	41 (34-52)	42 (39-50)	0.00
Mental functioning, Median (IQR)	54 (49-57)	42 (33-49)	30 (25-36)	26 (21-34)	21 (17-28)	0.00
Sick days, median (IQR)	0 (0-4)	1 (0-7)	6 (0-15)	7 (2-20)	8 (1-15)	0.00
Consultations in previous 12 months, Median (IQR)	9 (5-15)	11 (6-17)	14 (8-22)	16 (7-27)	12 (9-25)	0.00

[1]IQR – Interquartile range
[2] P-values from Kruskal-Wallis tests

The area under the ROC curve of the PHQ-9 was 0.87 (95% CI: 0.84-0.90). Table 3 shows the accuracy measures for different thresholds of the PHQ-9 score. The commonly used threshold of 10 had a specificity of 0.95 but a sensitivity of only 0.49. At a threshold of 6, sensitivity was 0.82 and specificity was 0.82. At this threshold the a-priori probability (prevalence) of 13% was increased to a posterior probability of 41% .

The categorical algorithm of the PHQ-9 showed a specificity of 0.98 and sensitivity of only 0.28. Based on this we defined an adjusted categorical algorithm to include the responses 'several days' as symptom present (see methods), whereas the original algorithm codes these answers as symptom absent. This resulted in a sensitivity of 0.84 and specificity of 0.81, close to those found for a threshold of 6 (Table 3). As the time delay between the PHQ-9 and the reference test varied, we performed an additional analysis to determine the influence of this time delay. Discrimination (area under the ROC curve) was similar when the delay between PHQ-9 and CIDI was longer (results not shown).

The area under the ROC curve of the PHQ-2 was 0.83 (95% CI 0.80-0.87). The commonly used threshold for the PHQ-2 of 3 showed a specificity of 0.94 and sensitivity of 0.42 (Table 3). As with the PHQ-9, lower thresholds showed more balanced values of sensitivity and specificity, notably at a threshold of 2. At this threshold, the a-priori probability (prevalence) of 13% was increased to a posterior probability of 34%.

Table 3 Sensitivity, specificity, and predictive values for different thresholds of the PHQ-9 and the PHQ-2. For details see text.

PHQ-9 threshold	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
≥4	0.89 (0.84-0.94)	0.64 (0.61-0.67)	0.27 (0.23-0.31)	0.97 (0.96-0.98)
≥5	0.86 (0.81-0.91)	0.75 (0.73-0.77)	0.34 (0.30-0.38)	0.97 (0.96-0.98)
≥6	0.82 (0.76-0.88)	0.82 (0.80-0.84)	0.41 (0.36-0.46)	0.97 (0.96-0.98)
≥7	0.74 (0.68-0.80)	0.87 (0.85-0.89)	0.47 (0.41-0.53)	0.96 (0.95-0.97)
≥8	0.65 (0.58-0.72)	0.90 (0.88-0.92)	0.51 (0.44-0.58)	0.94 (0.93-0.95)
≥9	0.57 (0.50-0.64)	0.93 (0.92-0.94)	0.55 (0.48-0.62)	0.94 (0.93-0.95)
≥10	0.49 (0.42-0.56)	0.95 (0.94-0.96)	0.59 (0.51-0.67)	0.93 (0.92-0.94)
≥11	0.44 (0.37-0.51)	0.96 (0.95-0.97)	0.61 (0.53-0.61)	0.92 (0.90-0.94)
≥12	0.39 (0.32-0.46)	0.97 (0.96-0.98)	0.64 (0.55-0.74)	0.92 (0.89-0.93)
PHQ-9 algorithm	0.28 (0.21-0.35)	0.98 (0.97-0.99)	0.69 (0.58-0.80)	0.90 (0.88-0.92)
PHQ-9 adjusted algorithm	0.84 (0.79-0.89)	0.81 (0.79-0.83)	0.40 (0.35-0.45)	0.97 (0.96-0.98)
PHQ-2 threshold				
≥1	0.90 (0.86-0.94)	0.58 (0.55-0.61)	0.24 (0.21-0.27)	0.98 (0.97-0.99)
≥2	0.81 (0.75-0.84)	0.76 (0.74-0.78)	0.34 (0.29-0.39)	0.96 (0.95-0.97)
≥3	0.42 (0.35-0.49)	0.94 (0.93-0.95)	0.53 (0.45-0.61)	0.91 (0.89-0.93)
≥4	0.31 (0.24-0.38)	0.97 (0.96-0.98)	0.64 (0.54-0.74)	0.90 (0.88-0.92)
≥5	0.19 (0.13-0.25)	0.99 (0.98-1.00)	0.69 (0.56-0.82)	0.89 (0.87-0.91)
≥6	0.14 (0.09-0.19)	0.99 (0.98-1.00)	0.67 (0.52-0.82)	0.88 (0.86-0.90)

PPV: Positive predictive value
NPV: Negative predictive value
95% CI: 95% Confidence Interval

Discussion

The PHQ-9 showed a very good internal consistency and test-retest reliability. Moreover, more severe depressive symptoms as measured by the PHQ were associated with poorer functional status, sick days, and higher number of general practice consultations. The accuracy of detecting major depressive disorder at recommended threshold of 10 and for the categorical algorithm, however, was poor. Lowering the threshold and minor adjustments of the categorical algorithm showed a considerable improvement of sensitivity, at the cost of lower specificity (Table 3). The adjusted categorical algorithm included all responses other than 'Not at all' as item present. The PHQ-2 showed a similar level of accuracy (i.e. sensitivity and specificity) when a lower threshold of 2 rather than 3 was used.

Our results of the reliability and construct validity of the PHQ-9 are similar to those reported in another primary care study[3] and a study of chronically ill primary care patients[13]. When we compared our observed sensitivities and specificities with other studies, we noted mixed results in the existing literature. A systematic review of the PHQ-9 in primary care found a pooled sensitivity of 0.77 (95% CI: 0.71-0.84) and a pooled specificity of 0.94 (95% CI: 0.90-0.97) for the diagnostic algorithm[15]. Similar results were found for the threshold of 10 in a systematic review by Gilbody et. al.[12]. Both reviews report substantially higher sensitivities compared to those reported here. However, a number of other studies in specific patients populations (e.g. patients with cardiovascular diseases) also observed low sensitivities and comparable specificities as we observed[13,17,27]. Similarly, the recommended threshold of 3 for the PHQ-2 showed a low sensitivity in comparison with other primary care studies[7,10,28], whereas other studies describe results similar to those reported here[16,17,29].

Strengths of this study are, first, that patients were included consecutively on random days, irrespective of their presented symptoms or signs and thus representing all patients in the waiting room of the GP. Second, patients were approached for participation in several general practices in both rural and urban areas to ensure a representative sample. Third, the reference test was administered by well-trained CIDI interviewers to guarantee the validity of the diagnoses and was applied in all attendees so that there was no selection bias. Fourth, this is the first study that validates the PHQ-9 and PHQ-2 in Dutch primary care.

Our study also has some limitations. First, the non-response rate for this study was relatively high. However, we found very minor non-significant differences in distributions of gender and age compared to responders (data not shown). Second, the prevalence of major depressive disorder in this study was relatively high[30-32]. It is possible that patients with major depressive disorder or similar mood problems were more willing to participate in our study. As a result, we would expect sensitivities and positive predicted values to be overestimated and specificities and negative predictive values to be underestimated[33]. This, however, is not consistent with the results presented here and therefore unlikely to explain our findings. Third, the test-retest reliability was

assessed in only 31 patients. Still, the results were very similar to earlier findings[3,13]. Fourth, the questionnaire was filled out at home. It is therefore possible that the answers were influenced by others (e.g. family members). However, if this explained our findings, this influence had to be systematically in one direction for patients with major depressive disorder and more or less absent for all other patients to explain our findings, which is unlikely. Furthermore, there was a time delay between the PHQ-9 and the CIDI. However, in an additional analysis, we observed no influence of the time delay on sensitivities and specificities of the PHQ-9. Also, a substantial part of the CIDI interviews was administered by telephone. Previous studies, however, have shown that telephone interviews are valid for clinical assessment of depression[31,34]. It has been suggested that the CIDI underdetects major depressive disorder when compared to the SCID[31]. In larger clinical or epidemiological studies, however, it is not feasible to administer the SCID in all patients because this is a semi-structured interview that has to be administered by clinicians instead of trained lay-persons. Also, most critical evaluations of the CIDI were based on earlier versions than the version (2.1) used in our study[35].

The limitations of our study cannot, in our view, explain the low sensitivities for detecting major depressive disorder we observed. Differences between the PHQ-9 and reference tests such as the CIDI and the SCID, have been previously described[15]. The PHQ-9 is designed to inquire about symptoms of major depressive disorder in the past 2 weeks rather than the past 12 months (adapted to the past 6 months in our study) for the CIDI. Patients with symptoms of major depressive disorder in the past 6 months and less severe symptoms in the past 2 weeks will not be detected with the PHQ-9 or the PHQ-2. Conversely, patients reporting little or no symptoms in the CIDI interview will also report no symptoms on the PHQ-9. As such, this difference in time frame could very easily result in low sensitivities and high specificities for the PHQ-9 threshold and algorithm and the recommended threshold for the PHQ-2.

The currently recommended high thresholds will lead to large numbers of undetected depressions. Before applied in clinical practice, lower threshold values as considered in the present study should be evaluated in other studies with new patients and different settings. The high negative predictive value and a relative low positive predictive value at the lower threshold of 6 (Table 3) showed that exclusion of major depressive disorder is more feasible than inclusion. Even though the positive predicted value of 41% still represents a considerable increase of the a-priori probability of 13%, it also emphasizes the need for a further diagnostic work-up for major depressive disorder in patients with a high score on the PHQ-9.

Conclusion

In conclusion, the results presented here indicate that the PHQ-9 and the PHQ-2 are useful instruments to detect major depressive disorder in primary care. As the positive predictive value is still low, a high score needs to be followed by an additional diagnostic work-up. In addition, the PHQ-9 is a valid measurement of depression severity. For both scales, however, clinicians should be aware that current recommended thresholds could lead to under detection.

References

(1) Spitzer RL, Williams JB, Gibbon M, First MB: The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. Arch Gen Psychiatry 1992, 49: 624-629.

(2) Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J et al.: The Structured Clinical Interview for DSM-III-R (SCID). II. Multisite test-retest reliability. Arch Gen Psychiatry 1992, 49: 630-636.

(3) Kroenke K, Spitzer RL, Williams JB: The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001, 16: 606-613.

(4) Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV, III, Hahn SR et al.: Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. JAMA 1994, 272: 1749-1756.

(5) Adewuya AO, Ola BA, Afolabi OO: Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. J Affect Disord 2006, 96: 89-93.

(6) Gilbody S, Richards D, Barkham M: Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM. Br J Gen Pract 2007, 57: 650-652.

(7) Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S et al.: Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord 2004, 78: 131-140.

(8) Wulsin L, Somoza E, Heck J: The Feasibility of Using the Spanish PHQ-9 to Screen for Depression in Primary Care in Honduras. Prim Care Companion J Clin Psychiatry 2002, 4: 191-195.

(9) Spitzer RL, Kroenke K, Williams JB: Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999, 282: 1737-1744.

(10) Kroenke K, Spitzer RL, Williams JB: The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003, 41: 1284-1292.

(11) Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL: Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. Psychosom Med 2001, 63: 679-686.

(12) Gilbody S, Richards D, Brealey S, Hewitt C: Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med 2007, 22: 1596-1602.

(13) Lamers F, Jonkers CC, Bosma H, Penninx BW, Knottnerus JA, van Eijk JT: Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. J Clin Epidemiol 2008, 61: 679-687.

(14) Persoons P, Luyckx K, Desloovere C, Vandenberghe J, Fischler B: Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. Gen Hosp Psychiatry 2003, 25: 316-323.

(15) Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC: Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. Gen Hosp Psychiatry 2007, 29: 388-395.

(16) Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T et al.: Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med 2010, 8: 348-353.

(17) McManus D, Pipkin SS, Whooley MA: Screening for depression in patients with coronary heart disease (data from the Heart and Soul Study). Am J Cardiol 2005, 96: 1076-1081.

(18) King M, Weich S, Torres F, Svab I, Maaroos H, Neeleman J et al.: Prediction of depression in European general practice attendees: the PREDICT study. BMC Public Health 2006, 6: 6.

(19) King M, Walker C, Levy G, Bottomley C, Royston P, Weich S et al.: Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study. Arch Gen Psychiatry 2008, 65: 1368-1376.

(20) American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 4th Text Revision edn. Washington, D.C.: American Psychiatric Association; 2000.

(21) Ter Smitten RH, Smeets RMW, Van den Brink W.: Composite International Diagnostic Interview - computerized version 2.1: Dutch translation and adaptation. World Health Organisation: Geneva; 1997; 2007.

(22) Kosinsky M: Scoring the SF-12 Physical and Mental Health Summary Measures. Medical Outcomes Trust Bulletin 1997, 5: 3-4.

(23) Little RJA, Rubin DB: Statistical analysis with missing data. New York: Wiley; 1987.

(24) Vach W: Logistic regression with missing values in the covariates. New York: Springer; 1994.

(25) Donders AR, van der Heijden GJ, Stijnen T, Moons KG: Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006, 59: 1087-1091.
(26) Greenland S, Finkle WD: A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995, 142: 1255-1264.
(27) Picardi A, Adler DA, Abeni D, Chang H, Pasquini P, Rogers WH et al.: Screening for depressive disorders in patients with skin diseases: a comparison of three screeners. Acta Derm Venereol 2005, 85: 414-419.
(28) Li C, Friedman B, Conwell Y, Fiscella K: Validity of the Patient Health Questionnaire 2 (PHQ-2) in identifying major depression in older people. J Am Geriatr Soc 2007, 55: 596-602.
(29) Cutler CB, Legano LA, Dreyer BP, Fierman AH, Berkule SB, Lusskin SI et al.: Screening for maternal depression in a low education population using a two item questionnaire. Arch Womens Ment Health 2007, 10: 277-283.
(30) Bijl RV, De Graaf R, Ravelli A, Smit F, Vollebergh WA: Gender and age-specific first incidence of DSM-III-R psychiatric disorders in the general population. Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). Soc Psychiatry Psychiatr Epidemiol 2002, 37: 372-379.
(31) Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR et al.: The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003, 289: 3095-3105.
(32) Waraich P, Goldner EM, Somers JM, Hsu L: Prevalence and incidence studies of mood disorders: a systematic review of the literature. Can J Psychiatry 2004, 49: 124-138.
(33) Brenner H, Gefeller O: Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 1997, 16: 981-991.
(34) Sobin C, Weissman MM, Goldstein RB, Adams P, Wickramaratne P, Warner V et al.: Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. Psychiatric Genetics 1993, 3: 227-234.
(35) Kurdyak PA, Gnam WH: Small signal, big noise: performance of the CIDI depression module. Can J Psychiatry 2005, 50: 851-856.

Predicting postherpetic neuralgia in elderly primary care patients with herpes zoster: prospective prognostic study

Wim Opstelten, MD, PhD, general practitioner[1]
Nicolaas P.A. Zuithoff, MSc, statistician[1]
Gerrit A. van Essen, MD, PhD, general practitioner[1]
Anton M. van Loon, PhD, virologist[2]
Albert J.M. van Wijck, MD, PhD, anesthesiologist[3]
Cornelis J. Kalkman, MD, PhD, professor of anesthesiology[3]
Theo J.M. Verheij, MD, PhD, professor of general practice[1]
Karel G.M. Moons, PhD, professor of clinical epidemiology[1]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
[2] Eijkman Winkler Institute for Microbiology, Inflammation and Infectious Diseases, University Medical Center Utrecht, The Netherlands
[3] Pain Clinic, Department of Anaesthesiology, Division Perioperative Care and Emergency Medicine, University Medical Center Utrecht, The Netherlands

Published in: Pain, 2007 nov, 132 Suppl 1:S52-9.

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

Postherpetic neuralgia (PHN), the most frequent complication of herpes zoster (HZ), can negatively affect quality of life. Many patients develop severe physical, occupational, and social disabilities as a consequence of their unceasing pain. Because the effect of treatment is disappointing once the syndrome has developed, the importance of PHN-preventive strategies is widely recognized.

For the timely identification of HZ-patients who might benefit from such (future) preventive strategies, it is important to know which factors predict PHN occurrence to facilitate selection of the HZ-patients with a higher risk of developing PHN. Knowledge of these predictive factors may also help researchers understand PHN's natural history and pathogenesis and contribute to the development and evaluation of preventive interventions. Previous research identified (advanced) age, greater rash severity, and more severe acute pain as predictors of increased PHN risk[1-3]. Some other potential predictors (painful prodrome,[2,4,5] ophthalmic localization[6], presence of anxiety and depression[7], and serological[8]/virological[9] factors) have also been studied. However, the prognostic value of combinations of these predictors – based on multivariable prediction modeling - has not yet been assessed. Hence, there is limited evidence on which of the above mentioned variables are true or independent predictors of PHN.

The purpose of our study was to assess which of the potential predictors reported in the literature independently contribute to the prediction of persistent pain following HZ. We were especially interested in whether psychological determinants and serological/virological parameters enhance the predictive value of the easy-to-obtain predictors like age and severity of pain.

Methods

Study population

The study population comprised 598 patients who had been included in the recently published PINE study[10,11], assessing the effectiveness of a single epidural injection of steroids and local anesthetics during the acute phase of HZ in preventing PHN in elderly patients. Briefly, 300 general practitioners in different regions of The Netherlands recruited patients from September 2001 to February 2004. Inclusion criteria were HZ within 7 days after onset of the rash, dermatome below C6, age > 50 years, sufficient command of the Dutch language, and willingness to comply with the allocated treatment and follow-up measurements. Exclusion criteria were coagulation abnormalities including use of coumarin anticoagulants (salicylates were allowed), bacterial infection of the skin overlying the vertebra of the affected dermatome, allergy to methylprednisolone or bupivacaine, and known serious immunity disorders (e.g. AIDS). Patients randomized to the control group received the current standard treatment for HZ of analgesics as needed and antiviral medication if the rash had been present < 72 hours. The general practitioner was free to select either acyclovir (800 mg five times daily), famciclovir (500 mg three times daily),

or valaciclovir (1000 mg three times daily), each administered orally for 7 days. In addition, those patients randomized to the epidural injection group was given a single epidural injection of 80-mg slow-release methylprednisolone-acetate and 10-mg bupivacaine within one working day after inclusion.

Baseline measurements included demographics, prodromal pain duration, acute zoster-associated pain severity, rash severity and duration, and psychological factors. For the present study, finger-prick blood was also collected at baseline from a random subset of 218 consecutive patients.

After one month, two months and three months, we sent a (same) questionnaire to the patients, requesting them to quantify their average pain experienced in the last 24 hours using a visual analogue scale (VAS) ranging form "no pain" at 0 mm to "worst pain ever experienced" at 100 mm.

Outcome: presence of significant zoster-associated pain after three months

Controversy exists about the definition of PHN. Most authors define PHN as pain persisting beyond a specific interval after rash outbreak[12-14]. Others define it as pain persisting beyond a specified interval after rash healing[15]. A recent systematic research, published after the PINE study's design and inception, showed that VAS scores <30 (on a 0-100 scale) are not associated with significant decrements in quality of life or the ability to carry out activities of daily living and are therefore not considered representative of PHN[16]. The outcome in our analysis, therefore, was zoster-associated pain rated ≥30 on the VAS scale three months after inclusion in the study.

Candidate predictors

Based on previous studies[1,2,4-6] we selected a priori 14 candidate predictors of the syndrome: 7 easy-to-obtain and 7 psychological predictors.

Easy-to-obtain predictors

The seven easy-to-obtain predictors included age, gender, rash duration (in days) and severity (mild [0-20 vesicles], moderate [21-46 vesicles], severe [≥47 vesicles][17]) at inclusion, prodromal pain duration (in days), pain severity at inclusion (VAS range: 'no pain' [0 mm] to 'worst pain ever experienced' [100 mm]), and use of antiviral medication. Since the PINE trial indicated that an epidural injection with steroids and local anesthetics was not associated with a reduced PHN occurrence[11], this intervention was not included as a potential predictor in our analysis.

Psychological predictors

The seven psychological predictors comprised five scores from the Pain Cognition List (PCL; I: negative self-efficacy score, II: pain catastrophizing score, III: positive expectation score, IV: resignation score, and V: trust in healthcare score; all ranging from 0 to 100)[18] and two predictors from the Dutch version of Spielberger's State-Trait Anxiety Inventory (STAI)[19]. The STAI

questionnaire is a validated questionnaire comprising 40 items with a four-point rating scale that measures anxiety state and anxiety disposition. Each score comprised 20 items: minimum and maximum values per score were 20 and 80, respectively.

Serological and virological predictors

To investigate whether varicella-zoster virus (VZV)-IgM, VZV-IgA, and VZV-IgG antibody titers and VZV viremia occurrence were additional PHN-predictors - i.e. beyond the above mentioned predictors - we analyzed finger-prick blood from a random subset of 218 patients at inclusion, i.e. before the first dose of antiviral drug (if applicable) was taken by the patient. We specifically chose for an early as possible measurement of serological/virological parameters in order to assess risk factors which may timely identify patients at high risk of PHN before (irreversible) damage has been inflicted and prevention has become pointless. The blood was collected on filter paper. After drying at room temperature, dried blood spots were transferred to plastic bags and shipped from general practices to the Department of Virology, University Medical Center Utrecht (Utrecht, The Netherlands), where they were stored at 4°C for up to 18 months before testing.

VZV antibodies (VZV-IgM, VZV-IgA, VZV-IgG) were determined using a commercial enzyme immunoassay. IgM and IgA assay results were expressed as the ratio between net optical density (OD) of a sample ($OD_{antigen} - OD_{control\ antigen}$) and 0.2 (OD cutoff value). IgG antibody titers to VZV were calculated according to the α-method, as recommended by the manufacturer, and transformed to a logarithmic scale.

Molecular detection of VZV-DNA was conducted using automated nucleic acid extraction together with an internally controlled real-time PCR assay, essentially as described by Stranska *et al.*[20].

Data analysis

We first estimated the (univariable) association between each predictor and the outcome. Then, without any univariable preselection, we used multivariable logistic regression modeling to assess which predictors independently contribute to prediction of PHN and to what extent using odds ratios with 95% CI. As we aimed to study whether more burdensome to measure predictors (i.e. psychological, serological, and virological factors) have added predictive value, we used a hierarchical modeling approach in which the simple predictors were included first[21,22]. The initial model with the 7 simple predictors was reduced by deleting (one-by-one) predictors with *P* values >0.15 based on the log-likelihood ratio test[23]. In contrast to etiological research, it is common and even recommended to use more liberal *P* values (i.e., >0.05) in prediction research[22-24]. The reduced model was then extended by adding, separately and in combination, psychological predictors to estimate their additional predictive value. Finally, in the subset of 218 patients, the best model to emerge from the above procedure was extended by adding, separately and in combination, the serological and virological predictors to establish their additional predictive value. Continuous

predictors were analyzed as linear terms as there were no indications of non-linearity based on cubic spline analysis[23].

The predictive accuracy of the reduced and extended prediction models was estimated by their calibration (reliability) and discrimination. Calibration was evaluated by graphically comparing the model's predicted probability of PHN with the observed proportions and tested with the Hosmer and Lemeshow test (H-L test)[25]. The model's ability to discriminate between patients with and without PHN was estimated using the area under the receiver-operating-characteristic curve (ROC area). In constructing this ROC curve, the multivariable model can in fact be considered a 'single' continuous predictor or test, existing of several component tests, with the model's estimated probability of PHN presence (ranging from 0 to 1) as the 'single' test result[23]. By estimating the model's sensitivity and 1-specificity at each probability threshold one can draw the ROC curve in a similar way as for a single predictor. Difference in ROC area between (reduced and extended) models was estimated accounting for the correlation between the models since they were based on the same subjects[26]. Generally, a model's predictive accuracy is too optimistic when its predictive performance is tested on the same data set from which it was developed, so-called over-fitting[23,24,27]. We thus used bootstrapping techniques, repeating the entire modelling process (including the variable selection process), to validate the final model, to adjust (shrink) the ROC area for over-fitting, and to obtain a shrinkage factor for the regression coefficients (log odds ratios) of the selected predictors[23]. A model's performance after bootstrapping better reflects its performance expected in future patients.

A total of 127 patients had values missing for one or more of the variables. The average percentage of missings per predictor was 5%. Since data omission rarely occurs at random, excluding subjects with missing values not only leads to loss of statistical power, but also to biased results[23,28]. To decrease bias and increase statistical efficiency it is better to impute missing values rather than perform complete case analyses. Accordingly, we imputed missing data using the linear-regression method (with addition of random-error term) available in SPSS software (version 12.0) before executing the above analysis.

Results

Total sample (598 patients)

Table 1 shows the baseline characteristics of the entire patient sample. Age and intensity of pain of the included patients were similar to the age and pain intensity of the HZ patients who met the inclusion criteria but were not included (data not shown). Of the 598 patients, 46 had PHN three months after inclusion (incidence: 7.7%) with a mean VAS score of 54.0 (SD 19.6).

Table 1. Univariable association between each candidate predictor and the presence or absence of postherpetic neuralgia (PHN) three months after the onset of herpes zoster (n=598)

Potential predictors	*All patients (n=598)*	*PHN[a] (n=46)*	*No PHN[a] (n=552)*	*OR[b]*	*(95% CI[c])*	*P[d]*
Age (years)	66.2 (9.8)	73.6 (7.1)	65.5 (9.8)	1.090	(1.054-1.127)	*< 0.001*
Female gender	60.9	58.7	61.1	0.907	(0.492-1.671)	*0.753*
Duration of prodromal pain (days)	5.3 (4.2)	4.4 (2.5)	5.4 (4.3)	0.925	(0.834-1.024)	*0.134*
Severity of pain at inclusion using VAS (mm)[e]	48.5 (26.6)	57.7 (30.3)	47.7 (26.1)	1.015	(1.003-1.027)	*0.015*
Duration of rash prior to inclusion (in days)	2.6 (1.7)	2.0 (1.4)	2.7 (1.7)	0.782	(0.642-0.952)	*0.014*
Severity of rash						
mild (0-20 vesicles)[f]	277 (46.3)	14 (30.4)	263 (47.6)	-	-	
moderate (21-46 vesicles)	201 (33.6)	15 (32.6)	186 (33.7)	1.515	(0.714-3.214)	*0.279*
severe (47 or more vesicles)	120 (20.1)	17 (37.0)	103 (18.7)	3.101	(1.475-6.519)	*0.003*
Antiviral medication	61.9	73.9	60.9	1.821	(0.923-3.595)	*0.084*
Questionnaires						
PCL factor I	37.6 (22.2)	37.4 (20.3)	37.6 (22.4)	0.999	(0.986-1.013)	*0.942*
PCL factor II	28.9 (23.6)	28.3 (22.6)	28.9 (23.7)	0.999	(0.986-1.012)	*0.853*
PCL factor III	41.4 (26.0)	43.5 (29.4)	41.3 (25.7)	1.003	(0.992-1.015)	*0.579*
PCL factor IV	48.2 (27.9)	47.2 (25.4)	48.3 (28.2)	0.999	(0.988-1.009)	*0.806*
PCL factor V	57.3 (25.2)	65.7 (20.8)	56.7 (25.4)	1.015	(1.002-1.028)	*0.021*
Spielbergers' STATE anxiety level	39.8 (10.2)	38.3 (8.5)	39.9 (10.3)	0.994	(0.963-1.025)	*0.686*
Spielbergers' TRAIT anxiety level	35.4 (9.9)	34.8 (8.0)	35.4 (10.1)	0.984	(0.954-1.014)	*0.295*

[a]Mean value (SD) or percentage; [b]Odds ratio; [c]95% confidence interval; [d]*P* value; [e]Visual Analogue Scale; [f]Reference category.

Advanced age, severe pain at inclusion, severe rash, antiviral medication, and a high PCL factor-V score were associated ($P<0.15$) with a higher PHN incidence, whereas longer prodromal pain duration and longer rash duration before consultation were associated with a lower PHN incidence (Table 1). Because the categories 'mild' and 'moderate' rash severity yielded similar associations with PHN occurrence, they were combined in subsequent analyses.

The overall basic multivariable model, including the seven 'easy-to-obtain' predictors, had an ROC area of 0.80 (before bootstrapping) and showed good calibration (H-L test: $P=0.49$). The reduced model included age, acute pain severity, rash duration before consultation, and rash severity and had an ROC area of 0.79 (H-L test: $P=0.27$). Antiviral therapy was not an independent PHN predictor in this model (odds ratio (OR): 1.38; $P=0.72$).

None of the two STAI questionnaire scores emerged as additional predictors of PHN after being added, separately or in combination, to the reduced model. When the five PCL factors were added, only PCL factor V ('trust in healthcare') was significantly associated with PHN. The ROC area increased significantly ($P=0.03$) to 0.81 (before bootstrapping). Hence, the statistically independent predictors were age, severity of pain at inclusion, rash severity, rash duration before consultation, and PCL factor-V score (H-L test: $P=0.76$; Table 2, left columns). After bootstrapping, the ROC area was 0.78 (optimism was 0.03).

Subset (218 patients): analysis of serological/virological determinants

Baseline characteristics of the 218 patients from whom blood samples were collected did not differ from the total sample of 598 patients (data not shown). None of the serological/virological variables were significantly associated with a higher PHN incidence in univariable analysis (Table 3). When adding these variables, separately or in combination, to the above final model derived from the total study population (Table 2), none were independently associated with PHN or increased the model's predictive accuracy.

Model for clinical practice

Although PCL factor-V score was significantly associated with PHN occurrence, it only slightly increased the ROC area. Because the burden of measuring this PCL factor in practice seems to outweigh its additional predictive ability, we decided to include age, rash duration prior to consultation, rash severity, and acute pain severity as independent predictors for use in daily practice (Table 2, right columns). The ROC area (after correction for over-optimism using bootstrapping techniques) of this practice model was 0.77 (95% CI: 0.71-0.82).

Table 2. The independent predictors of postherpetic neuralgia based on statistical testing (final model) and the predictors suggested for use in primary care

	Final model based on statistical testing			Predictors for use in primary care		
	OR[a]	95% CI[b]	*P*[c]	OR[a]	95% CI[b]	*P*[c]
Age (per year)	1.08	1.04-1.12	0.0022	1.08	1.04-1.12	0.0020
Duration of rash prior to consultation (days)	0.78	0.64-0.97	0.0109	0.78	0.63-0.96	0.0092
Severe rash (47 or more vesicles)	2.31	1.16-4.58	0.0061	2.34	1.18-4.62	0.0046
Severity acute pain (per VAS unit)	1.02	1.01-1.03	0.0049	1.02	1.01-1.03	0.0056
PCL factor V (per unit)	1.01	1.00-1.03	0.0316			
ROC area[a]	0.78	0.72-0.83		0.77	0.71-0.82	

[a]Odds ratio values are corrected for over-fitting by multiplying the estimated values with the shrinkage factor obtained from bootstrapping;
[b]95% confidence interval;
[c]*P* value.

Table 3. Univariable association between each serological and virological determinant and the presence or absence of postherpetic neuralgia (PHN) three months after the onset of herpes zoster in a random subset of patients (n=218)

Potential laboratory predictors	*PHN*[a] *(n=17)*	*No PHN*[a] *(n=201)*	*OR*[b]	*(95% CI*[c]*)*	*P*[d]
Serological determinants					
VZV[e]-IgM ratio at inclusion	0.4 (0.6)	0.5 (0.7)	0.87	(0.41-1.85)	*0.71*
VZV-IgA ratio at inclusion	3.7 (2.5)	4.0 (3.4)	0.97	(0.83-1.14)	*0.73*
log VZV-IgG at inclusion	6.7 (1.6)	7.0 (1.4)	0.87	(0.61-1.25)	*0.46*
Virological determinant					
VZV viremia at inclusion	5 (29.4)	39 (19.4)	1.73	(0.58-5.20)	*0.33*

[a]Mean value (SD) or number (%); [b]Odds ratio; [c]95% confidence interval; [d]*P* value; [e]Varicella-zoster virus.

Discussion

Our results show that older age, severe acute pain, and severe rash increase the risk of PHN development in elderly HZ patients, whereas longer rash duration before consultation reduces the risk of PHN. Although PCL factor-V-score was also an independent predictor, its marginal added value coupled with the burdensome measurement make it less useful in daily care. Scores on the other PCL factors and anxiety questionnaires did not predict PHN. Similarly, serological/virological variables did not independently contribute to a better prediction. Although these predictors were examined before, this is the first study to assess which are indeed the independent predictors of persistent zoster related pain. Because age and pain intensity were similar between included and non-included patients, the results of our study may be generalized to the defined target population.

This study confirms older age, greater acute pain severity, and greater rash severity as PHN-predictors[1,2,29] and they may now be considered as undisputed predictors. Contrary to some studies,[2,30,31] our results did not establish female sex as a predictor of PHN. Moreover, we were unable to assess ophthalmic HZ localization as a predictor[6,30,32,33] since we included only patients with HZ below the sixth cervical dermatome.

One unexpected finding was the independent predictive contribution of zoster rash duration before consultation: the longer the rash existed, the less often PHN occurred, independent of age, acute pain severity, and rash severity. This finding has not been previously reported, most likely because most studies included HZ patients with a rash duration <72 hours[2,5,34] or focused on the presence and duration of prodromal pain (which was no predictor in our analysis) rather than on zoster rash duration[4,33,35]. The predictive value of zoster rash duration may reflect patients' lack of concern, perhaps leading to both a restraint in consulting their physician and different perception of pain compared to patients who consulted their physician shortly after rash onset. The additional predictive value of PCL factor-V may also reflect this behavior: a higher score predicted PHN. PCL factor-V represents a patient's trust in healthcare, i.e. the expectation that others will find remedies to reduce their pain. Hence, people who underestimate their own contribution in recovery could be more prone to develop chronic pain. Our results support the view that psychological processes may contribute to PHN development[36,37]. In a small prospective study, Dworkin and colleagues focused attention on psychosocial PHN determinants and demonstrated that greater anxiety, greater depression, lower life satisfaction, and greater disease conviction at baseline predicted chronic zoster pain[7].

Our study showed that VZV-antibody titers have no predictive value with regard to PHN development. This contradicts the study by Higa *et al.*[8]. During their seven-week follow-up, the maximum antibody titers showed a positive linear relation with zoster pain duration. Apparently, antibody titers only influence the duration of acute and subacute zoster pain rather than chronic pain which we studied. We did also not establish a significant association between viremia at

baseline and PHN at three months. Scott *et al.*[9] recently showed that viremia at HZ presentation was significantly associated with zoster-associated pain ≥6 months later. Their analysis, however, was restricted to patients with virologically confirmed HZ, whereas our pragmatic study specifically investigated patients diagnosed using only clinical findings in accordance with daily (primary care) practice. Second, their blood-sampling method may be the reason for the higher percentage of VZV-DNA-positive patients (68% versus 20% in our study). We used PCR analysis on dried blood spots to facilitate virological analysis in primary care. Nevertheless, since highly sensitive VZV-DNA detection influences the test results of all HZ patients (with or without PHN), it is unlikely that the absence of the association between viremia and PHN in our study was completely caused by this difference in methods.

The prescription of antiviral medication did not independently predict PHN risk. The univariable odds ratio of 1.82 (0.92-3.60) might even suggest that these drugs increase PHN risk. This, however, was due to our design. Patients who received antiviral drugs essentially differed from those who did not with respect to rash duration before consultation: only HZ-patients who were included in the study within 72 hours after rash onset received antiviral drugs. Accordingly, rash duration, being a strong independent PHN-predictor in our study, possibly outweighed the predictive contribution of antiviral medication in multivariable analysis. Our design, therefore, does not allow inferences on the efficacy of antiviral drugs on PHN.

Our study has some potential limitations. First, physicians included HZ-patients based on clinical diagnosis, which was not confirmed serologically or virologically in most cases. Although general practitioners may have good clinical judgement with regard to HZ diagnosis[38,39], some cases may have been misclassified as HZ. Regardless, we aimed to identify independent PHN-predictors to serve general practitioners in their daily care, i.e. without the aid of laboratory investigation for HZ-diagnosis. Second, it would have been interesting to see the results of quantitative sensory testing (QST;[40]). Unfortunately, we did not corporate a standardized QST in our study protocol. Third, our definition of PHN remains debatable. We used the presence of significant (VAS>30) zoster-associated pain three months after rash onset. This definition was based on Coplan *et al.*[16] who showed that VAS scores <30 are not associated with significant decrements in quality of life and not considered to represent PHN. Since various other studies, including the PINE study[10,11] defined PHN as presence of zoster-associated pain one month after rash onset[13,14], we repeated our analysis using this PHN definition as endpoint. We found almost the same predictors. The only difference was that epidural injection, not age, was predictive. Epidural intervention was effective in reducing pain scores during the first month, but not thereafter[11]. The fact that age, the most reported predictor of chronic zoster-associated pain, was not predictive of pain during the first month, suggests that presence of zoster-associated pain after one month does not fully reflect chronic PHN. Fourth, the number of 46 patients with PHN was relatively low compared to the number of studied predictors[23,41]. This may have resulted in less stable estimates of the independent associations (odds ratios) of the predictors in the final model as well as of

the model's discriminative ability. Although we used internal validation techniques to adjust for this overfitting, we highly recommend that our findings, i.e. the estimated associations of the independent predictors and the predictive accuracy of our final model, are confirmed or validated in other HZ-populations[23]. Accordingly, we also did not construct a simple risk score for direct use in clinical practice: this should not yet be undertaken before the results of such validation studies have become available. The necessity for validation studies particularly applies to the role of serological/virological factors, as they were based on only 218 patients. However, as these serological and virological variables showed hardly any association in univariable or multivariable analysis, it is unlikely they will have predictive value in larger patient series. We confirmed this using an additional analysis in which we imputed the value of the serological/virological variables for the remaining 380 patients based on other characteristics observed in these subjects. Repeating the analysis after this imputation did not change our results. Similar odds ratios were found for the different serological/virological variables and none were significant. Finally, we recommend that our results also be validated in more comprehensive domains of HZ-patients, e.g., including cranial HZ-patients.

An effective therapy on the prevention of PHN is still lacking. Although many HZ-patients develop this pain syndrome despite antiviral therapy, the administration of antiviral drugs is one of the few interventions to shorten the duration of zoster-associated pain. Presently, several guidelines advice to prescribe these drugs according to age (e.g., above 50 or 60 years). The results of our study support a prescription which is not only based on age, but also on clinical findings (severity of acute pain and duration and severity of rash).

In conclusion, this large prospective study identified four simple predictors of PHN: age, severity of acute pain, severity of rash, and rash duration. These predictors may aid physicians in selecting high-risk HZ-patients who may benefit from preventive strategies, or at least should be monitored more closely in the acute period after rash onset.

Acknowledgements

The authors thank M. Schuller and N.M. de Vos, laboratory technicians, for their excellent help with the serological and virological analyses.

Financial support This study was funded by the Netherlands Organisation for Scientific Research (no. 945-02-009 and 917-46-360).

Conflicts of interest

There are no potential conflicts of interest in connection with this paper.

References

(1) Dworkin RH, Portenoy RK. Pain and its persistence in herpes zoster. Pain 1996:67:241-51.

(2) Jung BF, Johnson RW, Griffin DRJ, Dworkin RH. Risk factors for postherpetic neuralgia in patients with herpes zoster. Neurology 2004;62:1545-51.

(3) Coen PG, Scott F, Leedham-Green M, Nia T, Jamil A, Johnson RW, Breuer J. Predicting and preventing post-herpetic neuralgia: Are current risk factors useful in clinical practice? Eur J Pain 2006 [Epub ahead of print].

(4) Choo PW, Galil K, Donahue JG, Walker AM, Spiegelman D, Platt R. Risk factors for postherpetic neuralgia. Arch Intern Med 1997;157:1217-24.

(5) Whitley RJ, Shukla S, Crooks RJ. The identification of risk factors associated with persistent pain following herpes zoster. J Infect Dis 1998;178 Suppl 1:S71-5.

(6) Opstelten W, Mauritz JW, de Wit NJ, van Wijck AJ, Stalman WA, van Essen GA. Herpes zoster and postherpetic neuralgia: incidence and risk indicators using a general practice research database. Fam Pract 2002;19:471-5.

(7) Dworkin RH, Hartstein G, Rosner HL, Walther RR, Sweeney EW, Brand L. A high-risk method for studying psychosocial antecedents of chronic pain: the prospective investigation of herpes zoster. J Abnorm Psychol 1992;101:200-5.

(8) Higa K, Dan K, Manabe H, Noda B. Factors influencing the duration of treatment of acute herpetic pain with sympathetic nerve block: importance of severity of herpes zoster assessed by the maximum antibody titers to varicella-zoster virus in otherwise healthy patients. Pain 1988;32:147-57.

(9) Scott FT, Leedham-Green ME, Barrett-Muir WY, Hawrami K, Gallagher WJ, Johnson R, Breuer J. A study of shingles and the development of postherpetic neuralgia in East London. J Med Virol 2003;70:S24-30.

(10) Opstelten W, van Wijck AJ, van Essen GA, Buskens E, Bak AA, Kalkman CJ, Verheij TJ, Moons KGM. The PINE study: rationale and design of a randomised comparison of epidural injection of local anaesthetics and steroids versus care-as-usual to prevent postherpetic neuralgia in the elderly [ISRCTN32866390]. BMC Anesthesiology 2004;4.

(11) van Wijck AJ, Opstelten W, Moons KG, van Essen GA, Stolker RJ, Kalkman CJ, Verheij TJ. The PINE study of epidural steroids and local anaesthetics to prevent postherpetic neuralgia: a randomised controlled trial. Lancet 2006;367:219-24.

(12) Max MB, Schafer SC, Culnane M, Smoller B, Dubner R, Gracely RH. Amitriptyline, but not lorazepam, relieves postherpetic neuralgia. Neurology 1988;38:1427-32.

(13) Wood MJ, Balfour H, Beutner K, Bruxelle J, Fiddian P, Johnson R, Kay R, Cubed S, Portnoy J, Rentier B et al. How should zoster trials be conducted? J Antimicrob Chemother 1995;36:1089-101.

(14) Kost RG, Straus SE. Postherpetic neuralgia--pathogenesis, treatment, and prevention. N Engl J Med 1996;335:32-42.

(15) Tyring S, Barbarash RA, Nahlik JE, Cunningham A, Marley J, Heng M, Jones T, Rea T, Boon R, Saltzman R. Famciclovir for the treatment of acute herpes zoster: effects on acute disease and postherpetic neuralgia. A randomized, double-blind, placebo- controlled trial. Ann Intern Med 1995;123:89-96.

(16) Coplan PM, Schmader K, Nikas A, Chan IS, Choo P, Levin MJ, Johnson G, Bauer M, Williams HM, Kaplan KM, Guess HA, Oxman MN. Development of a measure of the burden of pain due to herpes zoster and postherpetic neuralgia for prevention trials: adaptation of the brief pain inventory. J Pain 2004;5:344-56.

(17) Whitley RJ, Weiss HL, Soong SJ, Gnann RW. Herpes zoster: risk categories for persistent pain. J Infect Dis 1999;179:9-15.

(18) Vlaeyen JWS, Geurts SM, van Eek H, Snijders AMJ, Schuerman JA, Groenman NH. De pijn cognitie lijst, experimentele versie (PCL-E) [The pain cognition list, experimental version (PCL-E)]. Swets & Zeitlinger, Lisse, The Netherlands, 1989.

(19) Spielberger CD, Gorush RL, Lushene RE. State Trait Anxiety Inventory manual. Consulting Psychologists Press, Palo Alto, CA, 1970.

(20) Stranska R, Schuurman R, de Vos M, van Loon AM. Routine use of a highly automated and internally controlled real-time PCR assay for the diagnosis of herpes simplex and varicella-zoster virus infections. J Clin Virol 2004;30:39-44.

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

R1 (22) Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337-8.

R2 (23) Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-87.

R3 (24) Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clin Chem 2004;50:473-6.

R4 (25) Hosmer D, Lemeshow S. Applied logistic regression. John Wiley and Sons, Inc., New York, 1989.

R5 (26) 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.

R6 (27) McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA 2000;284:79-84.

R7 (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.

R8 (29) Hope-Simpson RE. The nature of herpes zoster: a long-term study and a new hypothesis. Proc. of the Royal Society of Medicine 1965;58:9-20.

R9 (30) Hope-Simpson RE. Postherpetic neuralgia. J R Coll Gen Pract 1975;25:571-5.

R10 (31) Meister W, Neiss A, Gross G, Doerr HW, Hobel W, Malin JP, von Essen J, Reimann BY, Witke C, Wutzler P. A prognostic score for postherpetic neuralgia in ambulatory patients. Infection 1998;26:359-63.

R11 (32) Ragozzino MW, Melton LJ 3rd, Kurland LT, Chu CP, Perry HO. Population-based study of herpes zoster and its sequelae. Medicine (Baltimore) 1982;61:310-6.

R12 (33) Decroix J, Partsch H, Gonzalez R, Mobacken H, Goh CL, Walsh L, Shukla S, Naisbett B. Factors influencing pain outcome in herpes zoster: an observational study with valaciclovir. Valaciclovir International Zoster Assessment Group (VIZA). J Eur Acad Dermatol Venereol 2000;14:23-33.

R13 (34) Dworkin RH, Boon RJ, Griffin DR, Phung D. Postherpetic neuralgia: impact of famciclovir, age, rash severity, and acute pain in herpes zoster patients. J Infect Dis 1998;178 Suppl 1:S76-80.

R14 (35) Herr H. Prognostic factors of postherpetic neuralgia. J Korean Med Sci 2002;17:655-9.

R15 (36) Dworkin RH, Banks SM. A vulnerability-diathesis-stress model of chronic pain: herpes zoster and the development of postherpetic neuralgia. In: R.J. Gatchel and D.C. Turk (Eds.), Psychosocial factors in pain - critical perspectives. The Guilford Press, New York, 1999, pp. 247-69.

R16 (37) Livengood JM. The role of stress in the development of herpes zoster and postherpetic neuralgia. Curr Rev Pain 2000;4:7-11.

R17 (38) Kalman CM, Laskin OL. Herpes zoster and zosteriform herpes simplex virus infections in immunocompetent adults. Am J Med 1986;81:775-8.

R18 (39) Helgason S, Sigurdsson JA, Gudmundsson S. The clinical course of herpes zoster: a prospective study in primary care. Eur J Gen Pract 1996;2:12-6.

R19 (40) Rolke R, Baron R, Maier C, Tolle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Botefur IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihofner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 2006;123:231-43.

R20 (41) Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 1995;48:1495-501.

R21

R22

R23

R24

R25

R26

R27

R28

R29

R30

R31

R32

R33

R34

R35

R36

R37

R38

R39

5

Incidence and predictors of short and long term complications in pacemaker therapy: the FOLLOWPACE study

E.O. Udo, M.D. [1,2], N.P.A. Zuithoff [2], N.M. van Hemel, M.D.; PhD [1], C.C. de Cock, M.D.,PhD [3] ,T. Hendriks [3], P.A. Doevendans, M.D.; PhD [1], K.G.M. Moons, PhD [2]

[1] Department of Cardiology, UMC Utrecht, The Netherlands, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands

[2] Julius Center for Health Sciences and Primary Care, UMC Utrecht, The Netherlands, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[3] Department of Cardiology, VU Medical Center Amsterdam, The Netherlands, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands

Accepted for publication in Heart Rhythm

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

Since 1958 cardiac pacing has become the standard therapy for symptomatic and severe bradycardia due to brady-arrhythmia and conduction disorders. Worldwide implantation rates of pacemakers (PM) have strongly increased in the past years particularly in the elderly people[1].
Despite the impressive technologic development of implantable electronic cardiac devices and a wealth of clinical experience with their application, current pacing therapy appears not to escape from complications and technical failure. Multiple studies show that the majority of complications emerge shortly after implantation[2-6]. However, quantitative information about the type of complications and their incidence during long term follow-up is scarce, outdated or limited to specific device brands or patient populations.
The FOLLOWPACE study[7] started in 2003 in the Netherlands, and was designed to determine the incidence and predictors for short and long term complications after first PM implantation. The detailed successive follow-up of this prospective registry permits to identify patients at risk for adverse events. Knowledge about complications ameliorates the management of the PM recipient specifically regarding the standard in-hospital and modern trans-telephonic follow-up. This information also supports patient counselling before and after PM implantation. Moreover, because the FOLLOWPACE study was a multi-centre cohort without any pre-specified intervention or pacing therapy, the data serve as a benchmark for device clinics for comparing their complication frequency.

Methods

Patients

The FOLLOWPACE study is a prospective multicenter cohort study conducted in 23 PM centers in the Netherlands. The design of the FOLLOWPACE study has been published previously[7-11]. In brief, consecutive patients aged 18 years or older, who received a first PM for a conventional reason for chronic pacing[12], were eligible. Patients were not eligible if they were taking any investigational drug or had a non-approved or investigational PM implanted. In addition, patients with diseases at implant that are likely to cause death or severe morbidity during the 1st year after implantation such as active cancer were excluded. All patients provided written informed consent before PM implantation. This study complies with the Declaration of Helsinki and the protocol for this study was approved by the Ethical Commission of the University Medical Center Utrecht. Inclusion took place from January 2003 till November 2007. Follow-up lasted until 1 November 2010.

Potential predictors

At inclusion or the implantation visit (baseline), patient demographics, medical history and medication use were systematically recorded according to a pre-specified protocol. Furthermore, implantation procedure characteristics such as the indication for PM implantation, type of implanted PM and PM settings were also recorded, to quantify their potential, added predictive value for subsequent complications.

Finally, total annual PM implantations (including first implantations plus PM replacements) during the years of inclusion was scored for all centres (table 4). This predictor was used in further analysis as a proxy for hospital experience. It was studied for its additional predictive effects on the occurrence of short term complications beyond the patient and implantation related predictors. Due to the large variability of operators and fluctuations in the hospital staff during the inclusion period, experience of individual operators as a determinant for complications could not reliably be analysed.

Outcome measurements

After implantation, the frequency and intervals of subsequent follow-up visits were at the discretion of the medical professional in charge. At each (planned and unplanned) follow-up visit, technical and medical data, and the occurrence of a device or procedure related complication (PM complication) were systematically recorded. If a patient experienced multiple complications the clinical time course was reviewed to ensure that events counted were distinctly separate events. We divided complications in those occurring during the device therapy optimisation and lead maturation phase, i.e. within 2 months after PM implantation[13] (likely related to the implantation procedure; short term complications); and complications emerging thereafter during follow-up (long term complications). Patients who were lost to follow-up or died were censored, unless death was attributable to PM malfunction which was counted as a complication. As our goal was to determine the incidence and predictors for complications after first PM implantation, patients were censored at time of pulse generator replacement or upgrade.

Statistical analysis

The incidence of short and long term PM complications was estimated, and time to (or rather survival free from) any PM complication was estimated using the Kaplan-Meier method[14].

The prognostic relevance of the various baseline variables on the occurrence of both short and long term complications was assessed with Cox's proportional hazards regression models. Based on previous literature[3-5,15,16], we a priori selected the following twelve candidate predictors: age, gender, body mass index (BMI), indication for PM implantation, dual or single chamber device, prior cardiac surgery, prior cerebral vascular accident (CVA), the presence of coronary artery disease, cardiac valve disease, congestive heart failure, diabetes and hypertension. For predicting short term complications the pre-implantation use of anticoagulant drugs, route of

venous access, the manner of atrial and ventricular lead fixation (active or passive) and annual hospital implanting volume were added as potential predictors. For long term complications, the occurrence of a short term complication was studied as an additional predictor. For both outcomes, all candidate predictors were simultaneously included in the model. We used a p-value of <= 0.157 (based on the Akaike's Information criterion: AIC) to decide on retaining a variable as independent predictor in the model[17]. We estimated the discriminative ability of the final models using Harrel's c-index[18,19], which can be compared to the area under the receiver operating characteristic curve for a binary logistic model.

Information on all candidate predictors was complete, with the exception of BMI (missing in 6% of cases). As missing data are seldom missing completely at random, we followed methodological guidelines[20,21] and imputed these missing values using the single imputation method in SPSS (version 17).

Results

A total of 1517 patients were included in the FollowPace study and followed for a mean of 5.8 (SD 1.1) years, resulting in a total of 8797 patient years. Six patients were lost to follow-up: 4 because of data-loss after multiple hospital-transfers and 2 because of emigration abroad. Mean age at time of first implantation was 73.7 (±10.8) year and there were 856 (56%) males (table 1). Main indication for PM implantation was atrioventricular conduction disturbances in 613 (40.4%), sick sinus syndrome (SSS) or bradyarrhythmias in 557 (36.7%), atrial fibrillation with slow ventricular response in 266 (17.5%), and a PM was implanted in 81 patients (5.3%) for other indications (e.g. hypersensitive sinus carotis syndrome). Most implanted PM systems were dual chamber devices (68.8%).

The number of annually performed PM procedures ranged from 53 till 220 procedures/year (median 114, interquartile range 86-155). Participating centres constituted a mix of academic, teaching and non-teaching hospitals, with 7 centres performing 50-100 PM procedures/year, 9 centres performing 100-150 procedures/year and 7 centres >150 procedures/year (table 4).

Table 1. Baseline characteristics of all 1517 patients with a first pacemaker for conventional bradycardia indications.

	n	%
Male	856	56.4
Age [1]	73.7 (10.8)	
Body mass index [1]	26.3 (3.7)	
History		
Atrial tachy-arrhythmias	561	37.0
Cardiac surgery (CABG or valve surgery)	272	17.9
Coronary artery disease	301	19.8
Cardiac valve disease	326	21.5
Congestive heart failure	167	11.0
Prior cerebrovascular accident	159	10.5
Other cardiovascular disease	69	4.5
Diabetes	230	15.2
Hypertension	953	62.8
Use of anticoagulantia (ASA or coumarins)	942	62.1
Use of antiarrhythmic drugs	235	15.5
Main indication for implantation		
Atrio-ventricular conduction disturbances	613	40.4
Sick sinus syndrome, brady-tachycardias	557	36.7
Atrial fibrillation with slow ventricular response	266	17.5
Other	81	5.3
Implantation and PM related characteristics		
Vena subclavia used for venous access	1343	88.5
Vena cephalica used for venous access	174	11.5
Single chamber system AAI(R)	23	1.5
Single chamber system VVI(R)	381	25.1
Dual chamber system	1113	73.3
Passive atrial lead fixation	285	18.8
Passive ventricular lead fixation	1134	74.8
Pacing mode at discharge		
Dual	1043	68.8
Ventricular	401	26.4
Atrial	73	4.8

Data are presented as counts with percentages unless otherwise specified. [1] Mean with SD; CABG: coronary artery bypass grafting; ASA: acetylsalicylic acid.

Pacemaker related complications

Within two months after implantation there were 204 PM complications reported, occurring in 188 patients (12.4%), thereafter a total of 158 PM complications were reported in 140 patients (9.2%) (table 2). Figure 1 illustrates survival free from any PM complication. The curve shows a steep slope in the first six months followed by a more gradual decline, indicating a high incidence of PM complications early after implantation. At 1, 3 and 5 years, 15.6%, 18.3% and 19.7% of patients had suffered from a PM complication, respectively.

Traumatic complications - The most frequent occurring traumatic complication was a pneumothorax (2.2%), whereas damage to a cardiac structure was a relatively rare complication, occurring in 0.4% of implants. In one patient the tricuspid valve papillary muscle was damaged, for which no intervention was needed. Four patients (3 with a passive- and 1 with an active fixation RV-lead) experienced right ventricular wall perforation. In 3 cases this complications became apparent during the implantation procedure and did not lead to further complications, whereas in one patient this event became apparent after discharge due to cardiac tamponade, requesting uncomplicated pericardiocentesis. One patient in whom an active fixation atrial lead was implanted, experienced right atrial perforation that was treated by an uncomplicated pericardiocentesis. Only 1 late case of damage to a cardiac structure was reported during follow-up: an asymptomatic right ventricular wall perforation was coincidentally discovered by echocardiography.

Lead related complications - The most frequent lead related complication was dislodgement of a lead. Although mostly reported shortly after implantation, dislocation as a late problem is not uncommon. Within 2 months dislocation of an atrial lead occurred in 27 patients: in 16 (1.9%) patients with active atrial lead fixation and in 11 (3.9%) patients with passive atrial lead fixation (p=0.059). Dislocation of a RV-lead occurred in 20 (1.8%) passive fixation leads and in 4 (1.1%) active fixation leads (p=0.368). During follow-up dislocation occurred in 14 atrial leads (11 active (1.3%) vs 3 passive (1.1%), p=0.368) and in 10 RV-leads (3 active (0.8%) vs 7 passive (0.6%), p=0.661). The incidence of PM lead infection (i.e. lead endocarditis) was low (n=3; 0.2%).

Diaphragm or pocket stimulation was reported in 0.7% of patients, occurring equally often within 2 months as during follow-up. In most cases this disorder could be managed with output reprogramming, except for one patient in which repositioning of the lead was necessary because of frequent and uncontrollable complaints of diaphragm stimulation.

Other lead related problems reported within 2 months were as follows: in two procedures a right ventricular lead became constricted in the tricuspid valve apparatus and could not be removed. In both cases a second right ventricular lead was implanted without further complications, leaving the first lead in situ. In 3 procedures a lead was found to have too much tension in its curvature visualised by routine X-ray the day after implantation.

During follow-up after 2 months, 3 other lead related problems were reported: twice a slowly increasing lead impedance without rise in stimulation threshold, and once a high external noise signal on the EGM suggestive of impeding lead fracture.

Pocket complications - Pocket related complications were the second most often occurring type of complication. Although the incidence of difficult to control bleeding, necessitating reoperation was low (0.3%), conservatively managed haematomas are quite frequent, occurring in 2.9% of patients. Only 3 out of 45 patients with pocket hematomas, developed a pocket infection that all could be treated without reoperation.
Discomfort due to the pulse generator was not infrequent, occurring in almost 2%. In 11 (0.7%) patients this disorder necessitated a repeated surgical procedure.

Pulse generator problems - Problems with the pulse generator were seen in 5 patients within 2 months. A malfunction in the pulse generator connection screw for which another device was implanted in the same session, occurred in 4 patients and in one case a loose set screw was due to insufficient tightening.
During long term follow-up a PM recall was delivered in 11 patients: 4 devices were electively replaced, in 7 a 'watchful waiting policy' was applied, of these 2 were later replaced because of insufficient output. Four devices were unintentionally reprogrammed to their default settings, 2 of them after radiation therapy. Two devices were replaced because of lack of programmability for unknown reason. Four devices showed premature end-of-life caused by excessively and unnecessary high output settings chosen by an autocapture algorithm.

Table 2. Complications within 2 months and during long term follow-up occurring in 1517 patients with a first pacemaker.

	Within 2 months		During FU	
	n	%	n	%
Traumatic complications - total	**42**	**2,77**	**1**	**0,07**
Perforation of cardiac structure	6	0,40	1	0,07
Pneumo(hemo)thorax	34	2,24	0	0
Pericardial effusion	2	0,13	0	0
Lead related complications - total	**84**	**5,54**	**84**	**5,54**
Lead fracture †	2	0,13	6	0,40
Lead dislocation or disconnection †	50	3,30	24	1,58
Insulation problem †	4	0,26	11	0,73
Infection (i.e. lead endocarditis) †	0	0	3	0,20
Stimulation threshold problem	12	0,79	26	1,71
Diaphragm or pocket stimulation	11	0,73	10	0,66
Diaphragm or pocket stimulation †	0	0	1	0,07
Other *	5	0,33	3	0,20
Pocket complications - total	**72**	**4,75**	**49**	**3,23**
Haematoma	44	2,90	1	0,07
Difficult to control bleeding †	4	0,26	2	0,13
Infection	10	0,66	4	0,26
Infection †	4	0,26	8	0,53
Discomfort due to pocket or pacemaker	1	0,07	17	1,12
Discomfort due to pocket or pacemaker †	2	0,13	9	0,59
Skin erosion	7	0,46	8	0,53
Pulse generator problem	**5**	**0,33**	**23**	**1,52**
Problem with connection screw	5	0,33	0	0
Manufacturer recall	0	0	5	0,33
Manufacturer recall †	0	0	6	0,40
Reset to default settings	0	0	4	0,26
Device cannot be programmed	0	0	2	0,13
PM tachycardia	0	0	2	0,13
Malfunction of software algorithm	0	0	4	0,26
Total number of complications in need of reoperation	64	4,22	61	4,02
Number of patient experiencing a complication	188	12,4	140	9,20

* See text for details. † Complication is managed with reoperation. Numbers do not add up, because patients can experience multiple complications.

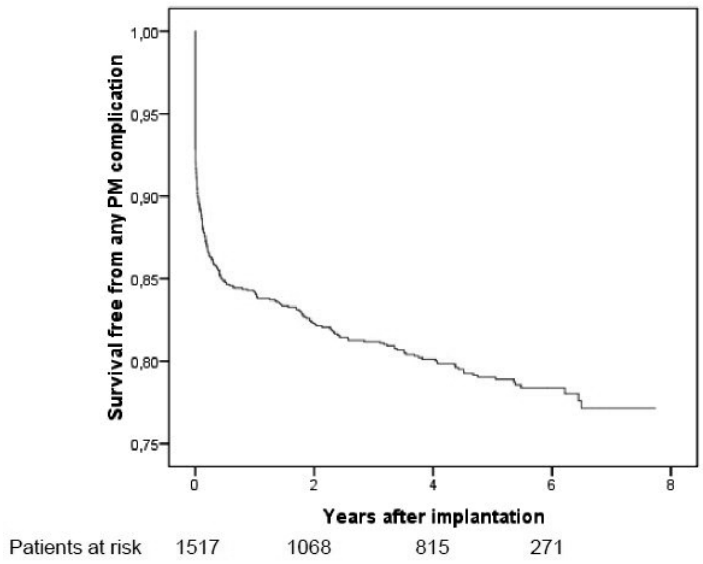


Figure 1: Kaplan Meier curve with survival free from any PM complication during a mean follow-up of 5.8 years.

Predictors for PM complications

Patient and Implantation related predictors for short and long term PM complications in univariable and multivariable analysis are shown in table 3. Independent predictors for a short term PM complication (i.e. occurring within 2 months) were male gender, age at implantation, body mass index, a history of cerebrovascular accident, congestive heart failure, use of anticoagulant drugs and passive atrial lead fixation. The c-index of this model was 0.62 (95% CI: 0.57-0.66), indicating rather poor discriminative ability between those who develop a complication from those who do not develop such event. Given these predictors, hospital implanting volume showed no independent predictive effect on the occurrence of PM complications (HR 1.0; p=0.78).

Independent predictors for long term PM complications were: age, body mass index, hypertension and a dual chamber device, yielding a c-index of 0.62 (95% CI: 0.57-0.67). A PM complication occurring within two months did not predict new PM complications.

Table 3. Relationship between patient and implantation characteristics and the occurrence of complications within 2 months and during follow-up in univariable and multivariable analysis.

	within 2 months				during follow-up			
	Univariable analysis		multivariable analysis		Univariable analysis		multivariable analysis	
Patient or implantation characteristic	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
Male gender	0.81 (0.61-1.08)	.16	0.72 (0.53-0.97)	.03	1.01 (0.73-1.42)	.94	0.91 (0.64-1.29)	.59
Age at implantation [1]	0.99 (0.97-1.00)	.02	0.98 (0.97-1.00)	.01	0.97 (0.96-0.99)	<.01	0.97 (0.96-0.99)	<.01
Body mass index [2]	0.59 (0.57-0.62)	.01	0.94 (0.90-0.98)	.01	0.97 (0.93-1.02)	.25	0.96 (0.92-1.01)	.09
Diabetes	1.11 (0.76-1.64)	.58	1.14 (0.76-1.69)	.53	1.18 (0.76-1.83)	.46	1.21 (0.76-1.90)	.42
Hypertension	1.00 (0.74-1.34)	.99	1.01 (0.72-1.41)	.96	1.12 (0.79-1.58)	.53	1.29 (0.87-1.90)	.20
History of atrial tachyarythmias	0.99 (0.73-1.33)	.94	1.00 (0.69-1.44)	.98	0.89 (0.63-1.26)	.50	0.87 (0.57-1.33)	.52
Prior cerebrovascular accident	1.60 (0.50-5.04)	.10	0.65 (0.37-1.14)	.13	1.13 (0.64-2.00)	.67	0.96 (0.54-1.71)	.89
Congestive heart failure	0.66 (0.45-0.99)	.04	1.36 (0.85-2.17)	.20	0.68 (0.43-1.09)	.11	1.24 (0.72-2.12)	.44
Cardiac valve disease	1.08 (0.77-1.52)	.65	1.03 (0.70-1.50)	.90	1.01 (0.67-1.51)	.98	1.06 (0.67-1.66)	.81
Coronary artery disease	1.06 (0.75-1.51)	.74	1.05 (0.71-1.55)	.80	1.13 (0.75-1.69)	.56	1.15 (0.74-1.79)	.52
Cardiac surgery	1.05 (0.73-1.51)	.80	1.06 (0.70-1.60)	.78	0.90 (0.58-1.41)	.64	0.84 (0.52-1.37)	.49
Main indication for implantation:		.39		.80		.64		.27
Atrio-ventricular conduction disturbances	ref		ref		ref		ref	
Atrial fibrillation with slow ventricular response	0.95 (0.62-1.44)		1.22 (0.71-2.10)		1.05 (0.66-1.66)		1.57 (0.87-2.86)	
Sick sinus syndrome, brady-tachycardias	0.98 (0.71-1.36)		0.96 (0.68-1.36)		0.87 (0.59-1.28)		0.89 (0.59-1.32)	
Other indications	1.56 (0.89-2.71)		1.13 (0.61-2.10)		1.32 (0.67-2.58)		0.91 (0.43-1.90)	
Dual chamber device	1.51 (1.05-2.16)	.03	3.09 (0.38-24.97)	.29	1.60 (1.05-2.45)	.03	1.67 (1.01-2.78)	.05
Vena subclavia used for venous access	0.92 (0.60-1.43)	.71	0.83 (0.53-1.29)	.40	-		-	
Use of aspirin or coumarin	1.13 (0.84-1.53)	.41	1.23 (0.87-1.74)	.24	-		-	
Passive atrial lead fixation	1.31 (0.93-1.85)	.13	1.36 (0.94-1.96)	.10	-		-	
Passive right ventricular lead fixation	0.99 (0.71-1.39)	.96	1.04 (0.73-1.50)	.82	-		-	
Annual hospital implanting volume [2]	1.00 (1.00-1.00)	.94	1.00 (1.00-1.00)	.78	-		-	
Implantation related complication	-		-		1.42 (0.83-2.43)	.20	1.28 (0.75-2.21)	.36

[1] Per one year increase in age; [2] Per unit increase; -: not studied for that outcome

Table 4. Characterisation of the participating hospitals in the FOLLOWPACE study.

Hospital	Hospital type	Cardiac care level *	EP lab	Included patients	Number of operators†	Mean number of PM procedures/year
1	teaching	1	no	12	4	146
2	teaching	1	no	160	9	111
3	teaching	2	no	109	5	220
4	academic	3	yes	32	3	100
5	general	1	no	38	2	118
6	general	1	no	163	4	114
7	teaching	3	yes	189	6	177
8	general	1	no	28	5	72
9	general	1	no	55	3	66
10	general	1	no	5	2	53
11	teaching	3	yes	126	6	186
12	academic	3	yes	7	4	154
13	teaching	1	no	6	6	155
14	teaching	1	no	20	3	86
15	teaching	1	no	17	4	203
16	general	1	no	16	3	59
17	teaching	1	no	16	3	80
18	teaching	1	no	101	7	125
19	teaching	2	no	4	1	126
20	general	1	no	60	5	113
21	teaching	2	no	41	5	162
22	general	1	no	250	6	114
23	general	1	no	61	4	85

EP lab: electrophysiology laboratory.

† The number of operators in an academic or cardiac teaching hospital will often include one or more trainees performing the implantation under close supervision of a staff cardiologist.
* Cardiac care level: 1) all standard cardiac care is provided including PM implantations; 2) In addition to standard cardiac care this hospital provides percutaneous coronary interventions. 3) In addition to standard cardiac care and percutaneous cardiac interventions, this hospital performs cardiac surgery.

Discussion

This large prospective Dutch cohort study on the incidence and predictors of PM complications in conventional pacing for bradyarrhythmias, observed PM complications within 2 months in 12.4% of 1517 patients. Whereas most studies report on complications that became apparent due to their need for a repeat surgical procedure, our large cohort study permitted the collection of all adverse PM events, in relation with other patient data. During long term follow-up for a mean of 5.8 (SD 1.1) years a PM complication was reported in 9.2% of patients. Patient and procedure related characteristics independently predictive of complications could be identified, but their discriminative ability is small.

Comparative studies

A valid comparison of complication rates is troubled by differences in definitions of complications resulting in a wide variance of outcomes, and secondly the strongly varying time windows of follow-up. In addition many previous studies present retrospective data and report about implantation procedures more than 20 years ago. Because of technical advancements, the reported numbers are not representative anymore for today's clinical practise. Furthermore, most studies only report on complications requesting repeated surgery[5,6,22,23]. Although these events are indeed major complications, many complications that impose a substantial burden to the patient do not require repeat surgery but should nonetheless be considered serious complications[16]. For this reason we restrict comparison to recent large scale studies published after 1995.

Early complications:

We report on complications within 2 months as this is a generally accepted time window in which complications directly related to a surgical procedure as PM implantation will have emerged[13]. The frequently used practise of reporting only in-hospital complications will surely underestimate true complication incidence. Only few studies extended their follow-up to adequately observe all adverse events[15].

A study of 1332 patients found an overall complication rate of 4.2% on acute complications (<48 hours after implantation)[24]. Parsonnet et al. reported 6 week complication rates ranging from 0.7% to 10.2%, with a mean overall rate of 5.7% for their 5 year study period in 632 patients[4]. This study is often referred to as an industry standard and it numbers are quite similar to ours, especially considering our slightly longer time window of two months. A recent study on the incidence of complications in PM therapy within 3 months found 69 complications requiring reoperation or prolonged hospitalization in 60 of 476 patients (12.6%)[2].

Considering the various sorts of complications, we found a low prevalence of heart perforation at implantation (0.4%), comparable to previous reports[2]. Pneumothorax was found in 2.2%. A single center study in a tertiary hospital in the UK including 1088 patients[3] reported pneumothorax in 1.8% of patients and a study in a university hospital in Finland found pneumothorax in 1.9% of 567 implanted devices[2].

Reoperation within two months was needed in 4.2% of patients in our series. A single center study of 1088 patients[3] found reoperation in 3.3% of patients. Reasons for reoperation included pocket infection in 0.9% (0.3% in our series), haematoma in 0.5% (0.3% in our series) and lead dislocation in 1.4%. Lead dislocation or disconnection in our series was found in 3.3% of patients. Recent studies found lead dislodgement in 3.7%[2] and over 3% of patients[5].

Pocket haematomas (not needing surgery) were not infrequent, occurring in 2.9% of patients in our series, compared to 3.2% during a 3 month follow-up period in 567 implanted devices[2].

One can conclude that although serious complications necessitating a reoperation are infrequent, the total number of patients experiencing an adverse event related to their PM implant (12.4%) is not insignificant.

Late complications

In the nineties several case reports for the first time showed the importance and risks of late PM complications[25]. We could only identify four studies reporting on PM complications occurring during long term follow-up[15,23,26,27]. Most of them are single center and retrospective[15,23,26], report only specific complications[15,23] and only complications in need of a repeat operation.

Eberhardt et al. report single center experience of 1884 patients with PM implantations between 1990 and 2001[23]. Only complications in need of reoperation were registered, and observed in 4.5% of this cohort. Late complications (after 3 months) occurred at an annual rate of 0.5%/year. Reason for late surgical revision was mostly lead problems.

The only multicentre data come from the MOST-study, in which the complication rate was 4.8% at 30 days, 5.5% at 90 days and 7.5% at 3 years[27]. The complication rates found in our study are higher, with 15.6%, 18.3% and 19.7% of patients suffering any PM complication at 1, 3 and 5 years, respectively. This is due to differences in definitions of complications, with 64% of complications in MOST being managed by reoperation, whereas in our study only the minority of complications needed reoperation.

A single center report on 446 PM implants showed a complication to occur in 14% of patients within a mean follow-up of 27 months[15]. Early complications (within 2 weeks after implantation) were seen in 6.7% of patients (4.9% were re-operated). Late complications developed in 7.2% of patients and reoperation was needed in 6.3%. Comparison to our extensive follow-up is not possible because of the very limited number of late complications surveyed in this report, as failure to capture or sense, clinical important AV-block in AAI-PM, PM infection and skin erosion.

Predictors for PM complications within 2 months

We found several patient characteristics to be independently predictive for short term PM complications. Males showed lower risk for complications than females (HR 0.72, p=0.30), which is in concordance with, amongst others, a recent large retrospective study by Nowak et al.[16,28].

The influence of ageing on complications has been a matter of debate and studies addressing this issue have shown contradictory results. Our observations show a lower incidence of complications in older patients. Older age was previously found to be associated with an increased risk of pneumothorax[16]. A study of 1,214 patients with AV-block showed age at implantation to be

an independent risk factor for complications[6]. The MOST-trial with 2,010 DDDR-implants for sinus node disease however could not show any relationship between age and complications[27]. Recently a retrospective analysis from a German quality control program was published[22]. This study of >17.000 implants found no increase in complications with increasing age with regard to the following complications (all requiring surgical intervention): pneumo/haemato-thorax, pericardial effusion, pocket haematoma, lead dislocation and device infection. Although this study had a large sample size, it only addresses a limited number of complications all requiring intervention and only during the in-hospital period, whereas longer follow-up data were not available. At present, there is little evidence to suggest higher follow-up complication rates in older patients.

Previously, a low BMI (<20) as well as a high BMI (>30) were shown to be predictive for cardiac perforation following PM implant[29]. Also lower weight was found predictive of pneumothorax[16] and skin erosion[15]. Our study confirms this relationship between lower BMI and complications, and observed a decrease by 9% (95% CI: 1-11%) in relative risk for every unit increase in BMI.

We found the use of anticoagulant drugs to be associated with a higher number of events. However information on INR-values or about the use of bridging therapy were not available. Furthermore, pocket haematoma or bleeding have been shown to occur more frequently in implantations by trainees than by experienced cardiologists[2], but our data do not permit to confirm this outcome.

We found passive atrial lead fixation to be correlated with a higher risk for complications. Active lead fixation may reduce the risk of early lead dislocation[30], but may be associated with a higher risk of cardiac perforation[25,27]. Neither the supposed lower dislocation rate with active fixation leads, nor an increased risk of cardiac perforation was seen in this study. Of note, we analysed passive lead fixation in relation to the composite endpoint of any complication, and not selectively atrial (or right ventricular) lead complications. Furthermore, predictors found in aetiological studies as ours, do not imply causality.

Hospital implanting volume did not have an independent predictive effect - beyond the independent patient and implantation related predictors - on the occurrence of short and long term complications. This finding is in concordance with a recently published study in which complication rates of small hospitals was found to be comparable to complication rates of larger hospitals[31]. Further, we do acknowledge that hospital volume may not necessarily reflect experience of the individual operator. However, our data did not allow for categorizing the experience of individual operators and therefore we applied hospital volume as proxy for hospital and operator experience. Furthermore, for more than 20 years, the Dutch Society of

Cardiology has implemented good clinical practise guidelines, which state a minimum of 40 first PM implantations/year/hospital. These guidelines intend to concentrate care of cardiac devices by specialised cardiologists. In all participating hospitals PM implantations were carried out by a select number of dedicated cardiologists, most of them exhibiting special interest and training in cardiac rhythm disorders.

Predictors for PM complications during follow-up

We found implantation related complications not to be predictive for future adverse PM events. To the best of our knowledge this relationship has not been studied before.

The relationship between dual chamber PM's and complications has been described before[6] and the increased risk is associated with the extra lead. Whether increasing age is associated with more frequent complications at implant remains disputable[22]. But the effects of ageing on complications during long term follow-up is fully unknown. Our data suggest no higher complication rate with increasing age. As well as being predictive for implantation related complications, lower BMI was associated with complications during long term follow-up.

Literature highlights that complications of PM implantation are largely determined by patient[23] and operator dependant characteristics[4,23,26]. In contrast, long term complications are mostly evoked by the device and leads. Because our study included many different types of devices and leads from all manufacturers resulting in subgroups too small to allow analysis, we could not explore the influence of specific devices and leads on complications.

Clinical implications

Our long term follow-up data are representative for current clinical practise as the study was fully embedded in routine care. The patients were recruited from all types of hospitals (academic, teaching and non-teaching) and were treated by operators and allied professionals with variable experience. This mixture of patient and technical characteristics suggests general applicability of the data that can have following clinical implications.

First, this study confirms that despite current better and easier implantable pacing devices and extensive tools for follow up, the complication rate and associated morbidity remains remarkable[32]. In efforts to diminish complication rates one could argue for concentration of PM implantations by limiting the number of implanting cardiologists, as complication rates have previously been shown to be related to operator experience[4,23,26]. Also, continuous education on technical and procedural features remains an indispensable measure to reduce the complication rate. Furthermore, when cost-effectiveness issues are being addressed, the additional costs associated with longer or repeated hospital stay, redo procedures with risk of new complications, and the burden for the patient, should be considered.

Secondly, we conclude that although several characteristics are independently related to the occurrence of PM complications, they were not capable of sufficiently identifying those at high risk of developing these complications. Identification of high risk patients would be useful for improving care to this population. Moreover, the ability to predict patients at risk could lead to a more tailored allocation of resources with varying the intensity of follow-up visits to the risk of the patient. Thereby decreasing the heavy workload associated with device follow-up, leading to more cost-effectiveness. Furthermore, these data could also be supportive in selecting patients suitable for remote monitoring.

Finally, these data can serve as a benchmark for other institutions to compare their local outcome and complication rate. The recording of future cohorts and comparing different case series can be part of an on-going cycle of quality of care control.

Limitations

These data result from a prospective, web based, nation-wide cohort study without any pre-specified intervention in the management of the PM patient. Complications were structurally assessed during each in-hospital follow-up visit for technical checks and patient examination, therefore the number of complications are meticulously registered. However the amount of burden to the patient was not quantified. Another limitation is our duration of follow-up, which had a maximum of almost eight years. Ideally all patients would be followed until death or replacement of the PM was required.

Conclusions

Despite technological advances in PM therapy, complication rates are still substantial. Although most complications occur in the initial post-implantation phase, complications during long term follow-up are not rare but include almost 10% of patients. Although various patient and procedure related characteristics are independently related to the occurrence of these PM complications, their ability to accurately identify patients at high risk is rather poor. Current guidelines on PM follow-up advise regular in-hospital follow-up for all patients[13,33]. Based on the data at hand, there is currently no evidence to change this practise.

References

(1) Schmidt B, Brunner M, Olschewski M, Hummel C, Faber TS, Grom A, et al. Pacemaker therapy in very elderly patients: long-term survival and prognostic parameters. Am Heart J 2003 Nov;146(5):908-913.

(2) Pakarinen S, Oikarinen L, Toivonen L. Short-term implantation-related complications of cardiac rhythm management device therapy: a retrospective single-centre 1-year survey. Europace 2010 Jan;12(1):103-108.

(3) Aggarwal RK, Connelly DT, Ray SG, Ball J, Charles RG. Early complications of permanent pacemaker implantation: no difference between dual and single chamber systems. Br Heart J 1995 Jun;73(6):571-575.

(4) Parsonnet V, Bernstein AD, Lindsay B. Pacemaker-implantation complication rates: an analysis of some contributing factors. J Am Coll Cardiol 1989;13(4):917-921.

(5) Moller M, Arnsbo P, Asklund M, Christensen PD, Gadsboll N, Svendsen JH, et al. Quality assessment of pacemaker implantations in Denmark. Europace 2002 Apr;4(2):107-112.

(6) Wiegand UK, Bode F, Bonnemeier H, Eberhard F, Schlei M, Peters W. Long-term complication rates in ventricular, single lead VDD, and dual chamber pacing. Pacing Clin Electrophysiol 2003 Oct;26(10):1961-1969.

(7) van Eck JW, van Hemel NM, Grobbee DE, Buskens E, Moons KG. FOLLOWPACE study: a prospective study on the cost-effectiveness of routine follow-up visits in patients with a pacemaker. Europace 2006 Jan;8(1):60-64.

(8) van Eck JW, van Hemel NM, de Voogt WG, Meeder JG, Spierenburg HA, Crommentuyn H, et al. Routine follow-up after pacemaker implantation: frequency, pacemaker programming and professionals in charge. Europace 2008 Jul;10(7):832-837.

(9) van Eck JW, van Hemel NM, Zuithof P, van Asseldonk JP, Voskuil TL, Grobbee DE, et al. Incidence and predictors of in-hospital events after first implantation of pacemakers. Europace 2007 Oct;9(10):884-889.

(10) van Eck JW, van Hemel NM, Kelder JC, van den Bos AA, Taks W, Grobbee DE, et al. Poor health-related quality of life of patients with indication for chronic cardiac pacemaker therapy. Pacing Clin Electrophysiol 2008 Apr;31(4):480-486.

(11) van Eck JW, van Hemel NM, van den Bos A, Taks W, Grobbee DE, Moons KG. Predictors of improved quality of life 1 year after pacemaker implantation. Am Heart J 2008 Sep;156(3):491-497.

(12) Gregoratos G, Abrams J, Epstein AE, Freedman RA, Hayes DL, Hlatky MA, et al. ACC/AHA/NASPE 2002 guideline update for implantation of cardiac pacemakers and antiarrhythmia devices: summary article. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/NASPE Committee to Update the 1998 Pacemaker Guidelines). J Cardiovasc Electrophysiol 2002 Nov;13(11):1183-1199.

(13) Wilkoff BL, Auricchio A, Brugada J, Cowie M, Ellenbogen KA, Gillis AM, et al. HRS/EHRA expert consensus on the monitoring of cardiovascular implantable electronic devices (CIEDs): description of techniques, indications, personnel, frequency and ethical considerations. Heart Rhythm 2008 Jun;5(6):907-925.

(14) Kaplan EL MP. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457-481.

(15) Kiviniemi MS, Pirnes MA, Ernen HJ, Kettunen RV, Hartikainen JE. Complications related to permanent pacemaker therapy. Pacing and clinical electrophysiology 1999;22(5):711-720.

(16) Link MS, Estes NA,3rd, Griffin JJ, Wang PJ, Maloney JD, Kirchhoffer JB, et al. Complications of dual chamber pacemaker implantation in the elderly. Pacemaker Selection in the Elderly (PASE) Investigators. J Interv Card Electrophysiol 1998 Jun;2(2):175-179.

(17) Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ 2009 Mar 31;338:b604.

(18) Harrell FE. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag; 2001.

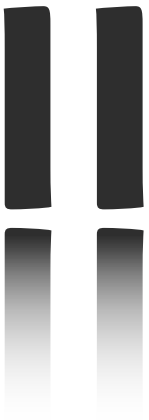
(19) Harrell FE,Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-387.

(20) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-1091.

(21) Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. : Wiley; 1987.

(22) Nowak B, Misselwitz B, Expert Committee Pacemaker, Institute of Quality Assurance Hessen. Effects of increasing age onto procedural parameters in pacemaker implantation: results of an obligatory external quality control program. Europace 2009 Jan;11(1):75-79.

(23) Eberhardt F, Bode F, Bonnemeier H, Boguschewski F, Schlei M, Peters W, et al. Long term complications in single and dual chamber pacing are influenced by surgical experience and patient morbidity. Heart 2005 Apr;91(4):500-506.
(24) Tobin K, Stewart J, Westveer D, Frumin H. Acute complications of permanent pacemaker implantation: their financial implication and relation to volume and operator experience. Am J Cardiol 2000 Mar 15;85(6):774-6, A9.
(25) Ellenbogen KA, Wood MA, Shepard RK. Delayed complications following pacemaker implantation. Pacing Clin Electrophysiol 2002 Aug;25(8):1155-1158.
(26) Harcombe AA, Newell SA, Ludman PF, Wistow TE, Sharples LD, Schofield PM, et al. Late complications following permanent pacemaker implantation or elective unit replacement. Heart 1998 Sep;80(3):240-244.
(27) Ellenbogen KA, Hellkamp AS, Wilkoff BL, Camunas JL, Love JC, Hadjis TA, et al. Complications arising after implantation of DDD pacemakers: the MOST experience. Am J Cardiol 2003 Sep 15;92(6):740-741.
(28) Nowak B, Misselwitz B, Expert committee 'Pacemaker', Institute of Quality Assurance Hessen, Erdogan A, Funck R, Irnich W, et al. Do gender differences exist in pacemaker implantation?--results of an obligatory external quality control program. Europace 2010 Feb;12(2):210-215.
(29) Mahapatra S, Bybee KA, Bunch TJ, Espinosa RE, Sinak LJ, McGoon MD, et al. Incidence and predictors of cardiac perforation after permanent pacemaker placement. Heart Rhythm 2005 Sep;2(9):907-911.
(30) Lemke B, Holtmann BJ, Selbach H, Barmeyer J. The atrial pacemaker: retrospective analysis of complications and life expectancy in patients with sinus node dysfunction. Int J Cardiol 1989 Feb;22(2):185-193.
(31) Haug B, Kjelsberg K, Lappegard KT. Pacemaker implantation in small hospitals: complication rates comparable to larger centres. Europace 2011 Nov;13(11):1580-1586.
(32) Maisel WH. Pacemaker and ICD generator reliability: meta-analysis of device registries. JAMA 2006 Apr 26;295(16):1929-1934.
(33) Vardas PE, Auricchio A, Blanc JJ, Daubert JC, Drexler H, Ector H, et al. Guidelines for cardiac pacing and cardiac resynchronization therapy: The Task Force for Cardiac Pacing and Cardiac Resynchronization Therapy of the European Society of Cardiology. Developed in collaboration with the European Heart Rhythm Association. Eur Heart J 2007 Sep;28(18):2256-2295.



Part II:
Methodological issues

6

Reporting of aims, designs, predictors, and outcomes in clinical prediction research: a systematic review

Nicolaas PA Zuithoff*[1], Walter Bouwmeester*[1], Susan Mallett[2,3], MI Geerlings[1], Yvonne Vergouwe[1, 4], Ewout W Steyerberg[4], Douglas G Altman[2], Karel GM. Moons[1**]

*equally contributed

[1] Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, the Netherlands

[2] Centre for Statistics in Medicine, University of Oxford, Oxford, UK

[3] Department of Primary Care, University of Oxford, Wolfson College Annexe, Oxford, UK

[4] Department of Public Health, Erasmus MC, Rotterdam, the Netherlands

Provisionally accepted in adapted form by PLoS Medicine

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

In recent years there has been an increasing interest in the methodology of prediction research[1–16]. Prediction research includes both diagnostic prediction studies studying the ability of variables or test results to predict the presence or absence of a certain diagnosis, and prognostic studies studying predictors of the future occurrence of outcomes[6,11,15]. Both types of prediction research may be studies of a single variable (or predictor or test) studies, multivariable studies aimed at finding the independently contributing predictors among multiple candidate predictors, or studies of the development, validation or impact assessment of multivariable prediction models. Many authors have stressed the importance of pre-defining the key aspects of a study including aims, study design, study population, clinically relevant outcomes, candidate predictors, sample size, and statistical analysis. Use of poor methods may lead to biased results[2,17–20].

We performed a comprehensive literature review of relevant articles published in high general medical journals to assess whether prediction research in the recent literature was conducted according to methodological recommendations. We focused on all types of clinical prediction studies and all (methodological) issues that are considered to be important in prediction research, rather than on specific types of outcomes (such as dichotomous outcomes[21]), specific methodological issues (such as missing data[22]), or specific disease areas (e.g. oncology)[19,20,23–25]. We studied the reporting and methods in clinical prediction research, focusing the specific study aims, study designs, study population, definition and measurement of outcomes and predictors, and the statistical power.

Methods

Literature search

We hand searched the 6 highest impact general medicine journals - The New England Journal of Medicine, Lancet, Journal of the American Medical Association, Annals of Internal Medicine, PLoS Medicine, and British Medical Journal - published in 2008. We excluded all studies that were not original research (e.g. editorials, letters) or had no abstract. One reviewer (WB) examined titles and abstracts of citations to identify prediction studies. The full text of all thus selected studies was obtained, and two authors (WB and NPAZ) independently assessed these studies for eligibility, and in case of doubt referred to a third independent reader (KGMM or YV).

Inclusion criteria

We included multivariable (>1 variable) prediction studies with one of the following aims: to find independently contributing predictors among multiple candidate predictors, to develop a prediction model, to validate or update an existing prediction model, or to quantify its impact on patient management or patient outcomes[6-9]. Prediction studies were defined as descriptive studies

where the aim was to predict an outcome by multiple (>1) independent variables, i.e. a causal relationship between independent variable(s) and outcome was not necessarily assumed[6,11]. We included both diagnostic and prognostic multivariable prediction studies. We excluded studies that investigated a single predictor, test or marker (such as single diagnostic test accuracy or prognostic marker studies), studies that studied only causality between one or more variables and an outcome, and studies that were not performed to directly contribute to patient care, for example finding predictors to predict citation counts.

Development of item list

We developed a comprehensive item list based on methodological recommendations for conducting and reporting of prediction research and discussions among the co-investigators. To this aim we studied existing reporting statements or checklists (i.e. CONSORT, REMARK, STARD, and STROBE) and an existing quality assessment tool (i.e. QUADAS) for aspects of study aims, design, and participant selection that also pertain to prediction studies[4,26–29]. Further, to identify additional aspects that are relevant for proper conduct and thus reporting of prediction research, we used publications on recommendations for conduct of prediction research and searched in their related articles and references in MEDLINE[1–10,12–16,19,20,30–34]. For external validation or model impact studies[8], separate items were defined and scored.

Data extraction

Data were extracted to enable a quantitative investigation of the good or bad reporting and methods influencing the quality of prediction studies. We extracted items regarding study aims, study design including participant or patient selection methods, study population, assessment and definition of outcomes and predictors, and statistical power (Box 1). Regarding the investigation of statistical power in prediction studies, we considered the reported multivariable models instead of the individual studies, because power differed among models within a single study.

Items were scored as present, absent, not applicable, or unclear. If an item concerned a numeric value (e.g. the number of patients), the scoring required a numeric value. If description was unclear, we counted it as not described or separately reported it in the tables. If included studies referred to other papers for detailed descriptions, the corresponding items were checked in those references.

Two authors (WB, NPAZ) independently extracted the data of the included studies. In case of doubt, items were discussed with a third and fourth reviewer (KGMM, YV). The inter-reviewer agreement of the data extraction was assessed by calculating the percentage of overall accordance between the two reviewers.

Box 1. Overview of review items addressed in this paper.

Study design	Type of clinical prediction study (e.g. model development), participant sampling or selection method (e.g. cohort or case-control approach)
Participants	Participant recruitment, follow-up, inclusion and exclusion criteria, setting (e.g. primary or secondary care or general population)
Predictors	Clear definition to ensure reproducibility, coding of predictor values, assessment blinded for outcome
Outcome	Clear definition to ensure reproducibility, type of outcome, assessment blinded for predictors
Statistical power	Effective sample size (e.g. number of outcome events compared to number of candidate predictors)

Analysis

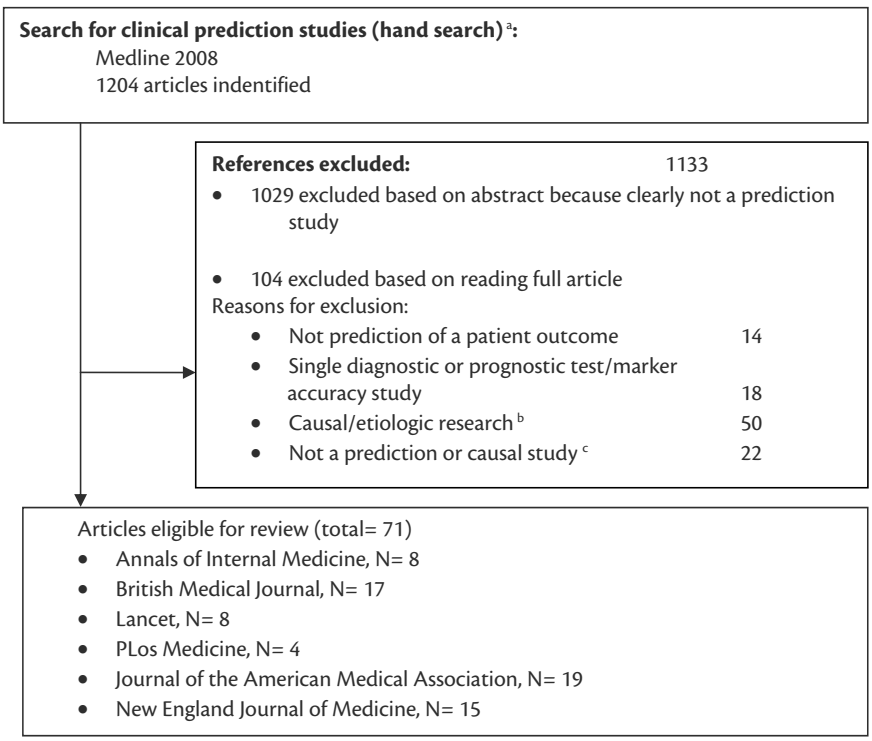
We grouped results by type of clinical prediction research, medical specialty (oncology, cardiovascular diseases, other) and whether the prediction analysis was a primary or secondary aim of the study. We distinguished 5 types of multivariable clinical prediction research:

- *Predictor finding studies* which aim to discover or explore which predictors or variables out of a number of candidate predictors, independently contribute to the prediction of, i.e. are associated with, an outcome[3,6,31].
- *Model development studies (without external validation)* which aim to develop a multivariable prediction model, e.g. for use in medical practice to guide patient management. Such studies aim to identify the important predictors, assign the (mutually adjusted) weights per predictor in a multivariable analysis, and develop a final multivariable prediction model. A key aspect is to estimate the model's predictive performance (e.g. calibration and discrimination statistics)[6,7].
- *Model development studies with external validation* which have the same aim as the previous type and also aim to test the performance of the developed model in a so-called external dataset from another time period (temporal validation) or other hospital, country or setting (geographical validation). Withholding the data for some centres for validation was also considered as (geographical) external validation, but random split sample methods were not[8].
- *External validation studies (with or without model updating)* which aimed to assess the performance of an existing prediction model using external patient data that were not used in the development process, and (possibly) adjusted or updated the model based on the validation data set[8,35].
- *Impact studies* which aim to quantify the effect or impact of using a prognostic or diagnostic prediction model on physicians' behaviour, patient outcome or cost-effectiveness of care relative to not using the model or usual care[9,35,36].

Results

We identified 1204 articles by hand searching, of which 71 met the inclusion criteria (figure 1 and appendix). Most studies were excluded based on title or abstract. During the search, it was hard to distinguish, only based on the abstract, between descriptive predictor finding studies and articles studying causality of a prognostic factor with an outcome. Therefore, we had to study the full texts, resulting in 50 causal prognostic studies excluded in the second selection round (flowchart).

Figure 1. Flowchart of included studies.



[a] The hand search included only studies with an abstract, published in 2008 in the New England Journal of Medicine, Lancet, the Journal of the American Medical Association, Annals of Internal Medicine, and PLoS Medicine. The following publication types were excluded beforehand: editorials, bibliographies, biographies, comments, dictionaries, directories, festschrifts, interviews, letters, news, and periodical indexes.
[b] Studies, generally conducted in a yet healthy population, aimed at quantifying a causal relationship between a particular determinant or risk factor and an outcome, adjusting for other risk factors (i.e. confounders).
[c] An example Pletcher MJ, et al. Prehypertension during young adulthood and coronary calcium later in life. Ann Intern Med. 2008 Jul 15;149(2):91-9.

Data extraction

The two reviewers agreed on a median of 92% (IQR 75%-100%) of the items extracted. Most discrepancies related to specific patient sampling strategies or patient sources in the reviewed articles and were resolved after discussion with a third independent reviewer.

A challenge was to objectively distinguish between predictor finding and model development studies. Authors did in general not explicitly state their aim, so we studied the full text to stratify the studies as predictor finding or model development study. This required interpretation of the reviewers.

Study aim

Most multivariable prediction studies were published in the field of cardiovascular diseases (n=24) (Table 1). The aim was mostly to identify independently contributing predictors of an outcome (n=51/71) (Table 1). Of the prediction modelling studies (N=20), the vast majority included model development studies without (n=11) or with (n=3) an external validation included. Pure external validation and model impact studies were rare (n=6). There were few multivariable diagnostic studies (n=5/71). In these 71 publications 135 models or sets of predictors were studied. For example, in predictor finding studies multivariable modelling to search for the independently contributing predictors was applied across different patient subgroups (e.g. males versus females), for multiple outcomes; and in prediction modelling studies more than 1 model was developed (for e.g. different outcomes or presenting a basic and extended model), validated or assessed for its impact.

Design of patient sampling

A cohort, nested case-control or case-cohort design are commonly recommended for prognostic and diagnostic model development and validation. A prospective cohort is preferable, because it enables optimal measurement of predictors and outcome. A retrospective cohort allows a longer follow-up period but usually at the expense of poorer data[6]. Randomized trial data has similar advantages as prospective cohort data, if eligibility criteria are not too restrictive leaving a study population which is not representative. Further, treatments proven to be effective in the trial should be included or adjusted for in the prediction model. Cohort, nested case-control, or case-cohort data allow calculation of absolute outcome risk[6,18,37]. A (non-nested) case-control design may, however, be sufficient for predictor finding studies since these studies generally do not aim to calculate absolute risks.

We found that case-control designs were indeed only used by predictor finding studies (Table 2). Prospective cohort data, either observational or (randomized) trial data, were most frequently used (n=44, 62%). Three cohort studies had a cross-sectional design, which was possible because predictor values did not change (gender, genes etc) or because it involved a diagnostic prediction study.

Table 1. Aim of the included multivariable prediction studies, subdivided by clinical domains. Numbers are column percentages, and absolute numbers between parenthesis.

	Cardiovascular (N=24)	Oncology (N=13)	Other [a] (N=34)	Total papers (N=71)	Number of models (N=135)	Number of diagnostic studies
Predictor finding studies						
Prediction was primary aim	46 (11)	62 (8)	44 (15)	48 (34)	49 (66)	1
Prediction was secondary aim	17 (4)	31 (4)	26 (9)	24 (17)	21 (28)	
Prediction model development without external validation	21 (5)	8 (1)	15 (5)	15 (11) [b]	14 (19)	1
Prediction model development with external validation	4 (1)	0 (0)	6 (2)	4 (3)	8 (11)	0
External validation, without updating a prediction model [c]	8 (2)	0 (0)	3 (1)	4 (3)	5 (7)	1
Impact assessment of a prediction model	4 (1)	0 (0)	6 (2)	4 (3)	3 (4)	2

[a] Including studies from infectious diseases (n= 7), diabetes(n=5), neonatology & child health (n=6), mental disorders (e.g. dementia) (n=4), or musculoskeletal disorders (e.g. low back pain) (n=4).
[b] Three of these model development studies also updated a previously developed model, without external validation. Reference [37, 45, 61] in the appendix.
[c] There were no external validation studies of a previously published model which also updated the model after poor validation.

Table 2. Study design in relation to study aim. Numbers are column percentages, and absolute numbers between parenthesis [a].

	Total (N=71)	Predictor finding studies (N=51)	Development without external validation (N=11)	Development with external validation (N=3)	External validation (without updating) (N=3)	Impact analysis (N=3)	Comments (N) [a]	
Prospective cohort [b]	62 (44)	53 (27)	82 (9 [c])	100 (3 [b])	67 (2)	100 (3)	Cross-sectional	1
							Randomized trial	13
Retrospective cohort	14 (10)	16 (8)	9 (1)	0 (0)	33 (1)	0 (0)	Cross-sectional	2
Case-control [d]	8(6)	12 (6)	0 (0)	0 (0)	0 (0)	0 (0)	Nested	4
							Non-nested	2
Not described or unclear	15 (11)	20 (10)	9 (1)	0 (0)	0 (0)	0 (0)		

[a] Numbers are percentages, and absolute numbers between parenthesis, except for the column 'Comments', which included only absolute numbers.
[b] Some prediction studies had a cross-sectional cohort design, which was possible because predictor values did not change (gender, genes etc) or because it involved a diagnostic prediction study.
[c] Of the 13 studies which used randomized trial data, 11 were predictor finding or model development studies. Of these 11 studies, 5 adjusted for the treatment effect, 3 did not because there was no treatment effect, 1 did not adjust despite an effective treatment, and in 2 studies such adjustment was missing.
[d] One study used 2 designs: a cross-sectional case-cohort and a nested case-control (here both scored as nested case-control)

The analysis of impact of a prediction model on patient outcome requires a comparative study design[9,36]. A randomized trial was used by two impact studies; the third used a before-after design (comparing patient outcomes before and after the introduction of a prediction model).

Participant recruitment, follow-up and setting

Participant recruitment was in general well described. Inclusion criteria were reported in 64/71 (90%) studies. Description of the cohort characteristic was clear in 68/69 of the relevant prediction studies (not applicable for 2 case-control studies). Recruitment dates were reported in 88% of the studies. Length of follow-up was not reported in 9 studies, making it impossible to understand to what time period the predicted risks of the predictors apply, limiting the applicability of study results. Whether (all) consecutive patients were included or how many participants refused to participate was hardly reported and could not be scored. The majority of studies included patients from the hospital setting (38%), or from the general (healthy) population (27%). Setting was not reported in 4% of the studies (Table 3).

Table 3. Reporting of inclusion criteria, participant recruitment and follow-up, and setting

	Percentage (N= 71)	Comments
Inclusion and exclusion criteria	90 (64)	
Recruitment dates reported		
Start	89 (63)	
End	87 (62)	
Follow-up dates reported		NA in 12 studies, because of a cross sectional design
Start	85 (50)	
End	95 (56)	
Setting		
Primary care	8 (6)	0 existing registries
Hospital care	38 (27)	3 existing (hospital) registries [a]
General population	27 (19)	3 existing registries
Combined primary and secondary care	6 (4)	1 existing registry
Combined general population and hospital care	11 (8)	1 existing registry
Other	6 (4)	4 existing registries
Unclear or not reported	4 (3)	3 existing registries

[a] existing registry defined as: routinely collected (medical) data that were not initially collected for research purposes

Outcome

In the outcome reporting, we expected differences between studies with prediction as primary or secondary aim, but this was not observed. Outcomes were well defined in 62/68 (91%) studies (Table 4). However, only 12 (22%) studies reported that they blinded the outcome measurement

for predictor values. Knowledge of the predictors might influence outcome assessment, resulting in a biased estimation of the predictor effects for the outcome[6,17,27]. In one study, it was clear that the predictor was also used to determine the presence of the outcome.

Most studied outcomes were binary (23/63; 34%) or time to event (30/63; 48%) outcomes (Table 4). Some outcomes are binary by necessity, however in some studies, continuous, categorical and time to event data were analysed as binary outcomes, a practice that is not recommended as less accurate predictions are likely to result with similar detriment to dichotomizing predictor variables [38]. Other outcomes were analyzed in two ways (e.g. a time-to-event outcome which was analyzed as time-to-event and as a binary outcome neglecting time), therefore the summation of 6 + 23 + 8 + 30 is higher than the total of 63 studies. If a study analyzed 2 binary outcomes, it was counted as 1 binary outcome in Table 4.

Table 4. Reporting of outcome[a]

		Percentage (N)
Clear definition		91 (62)
Assessment blinded for predictors [b]		22 (12)
Type of outcome described [c]		93 (63)
Continuous		9 (6)
	Linear regression	83 (5)
	Logistic regression [d]	17 (1)
Binary		34 (23)
	Logistic regression	91 (21)
	Non-regression [e]	9 (2)
Categorical		12 (8)
	Polytomous regression	38 (3)
	Logistic regression	50 (4)
	CART [f]	13 (1)
Time to event		48 (30)
	Survival analysis	97 (29)
	Logistic regression	3 (1)

[a] Impact studies were excluded for this table because these studies had outcomes of a different type (e.g. costs). Hence, the total number of studies is 68.
[b] NA in 11/68 studies, because all cause death was the outcome.
[c] Type of outcome and how they were analyzed (unclear for 5 studies). The summation of 6 + 23 + 8 + 30 is higher than 63, because some outcomes were analyzed in two ways (e.g. a time-to-event outcome which was analyzed as time-to-event and as a binary outcome neglecting time). If a study analyzed 2 binary outcomes, it was here counted as 1 binary outcome
[d] After dichotomization of a continuous outcome.
[e] One study used the Cochran–Mantel–Haenszel procedure, another calculated odds ratios.
[f] CART= classification and regression tree.

Table 5. Number of outcomes studied, by type of prediction study. All numbers are column percentages and absolute numbers between parenthesis.

Type of prediction study	Total (N=70) [a]	Predictor finding studies (N=50)	Development without external validation (N=11)	Development with external validation (N=3)	External validation without updating [b] (N=3)	Impact analysis [b] (N=3)
Number of outcomes studied						
1	60 (42)	62 (31)	73 (8)	33 (1)	67 (2)	0 (0)
2	26 (18)	22 (11)	18 (2)	67 (2)	33 (1)	67 (2)
≥ 3	14 (10)	16 (8)	9 (1)	0 (0)	0 (0)	33 (1)
Combined endpoint [c]	20 (14)	18 (9)	27 (3)	67 (2)	0 (0)	0 (0)

[a] Unclear for 1 study

[b] For external validation: the number of outcomes for which the accuracy of the prediction model was tested; for impact analysis: effect of prediction model on patient outcome, model consumer (physician) behavior or costs.

[c] It was a combined endpoint if 1) the authors explicitly stated that they analyzed a combined endpoint and 2) when the predicted outcome was both fatal and nonfatal events. All cause death was not considered a combined endpoint.

Prediction of more than 1 outcome was very common in predictor finding studies, apparently due to their exploratory aim (Table 5). However, selective reporting of outcomes (and predictors) could be a risk[39]. Unfortunately study registration is not mandated for prediction research, so it is generally impossible to assess whether some outcomes were analysed but not reported.
Some studies predicted a combined endpoint (14/71; 20%) (Table 5). The use of a combined endpoint will give problems if the predictor effect is in opposite direction for two outcomes included in the composite endpoint[40,41].

Predictors

Description of predictor variables was in general clear (59/68; 87%) (Table 6). In 51/68 (75%) of the studies, predictor measurement was blinded for the outcome, because of the prospective design. Only 7 non-prospective studies explicitly blinded predictor measurement for the outcome. One study also assessed the predictors independently, i.e. the new predictor was assessed without knowledge of the other predictors. Predictor interaction (non-additivity) was tested in only 25 of the 51 predictor finding studies, and in 11 of the 14 model development studies. It was somewhat unexpected that in only half of the predictor finding studies the effect of predictors across different groups or other predictors was tested, because of their explorative aim. Nevertheless, testing of all predictor interactions can lead to a multiple testing issues, which is discussed in the next section. Dichotomization of continuous predictors, which has been discouraged for decades[3,38], is still common practice (21/64; 33%).

Table 6. Reporting of candidate predictor [a]

	Percentage (N)
Clear definition	87 (59)
Assessment blinded for outcome(s)	75 (51)
Predictor part of outcome	1 (1)
Interaction of predictors tested [b]	55 (36)
Handling continuous predictors described [c]	67 (43)
Kept linear (continuous)	67 (43)
(Fractional) polynomial transformation or any spline transformation	19 (12)
Categorized	47 (30)
Dichotomized	33 (21)
Other	3 (2)

[a] Impact studies (N= 3) were excluded for this table as their aim is not to develop or validate a prediction model, but rather to quantify the effect or impact of using a prediction model on physicians' behaviour, patient outcome or cost-effectiveness of care relative to not using the model or usual care. Hence, total N=68.
[b] Not applicable for 3 external validation studies. Hence N=65.
[c] Not applicable in 4 studies, because one studied no continuous predictors, and 3 were external validation studies. Hence, N=64. Unclear in 19 studies, not described in 2 studies. The summation of 43 + 12 + 30 + 21 + 2 is higher than 43, because some studies handled continuous predictors in two ways (e.g. dichotomizing blood pressure, and categorizing body mass index in 4 categories).

Statistical power

For judgement of statistical power in studies estimating predictor effects, a rule of thumb is 10 events per candidate predictor[42–44]. For continuous outcomes, the effective sample size is determined by the number of patients in linear regression analysis. For categorical and binary outcomes, the number of patients in the smallest group determines the effective sample size. For time to event outcomes, statistical power is related to the number of patients who experience the event. The number of candidate predictors should include all variables considered in the analysis, the number of predictor interactions, and the number of dummy variables used to include a categorical variable in the model. We attempted to include all these variables, however studies rarely had sufficient reporting to enable clear assessment of the statistical power.

For prediction as primary aim and prediction as secondary aim studies, we calculated the effective sample size of the models based on the number of predictors in the final model and based on the number of candidate predictors (Table 7a). Statistical power was overestimated when only the number of predictors in the final model was counted. The number of events or candidate predictors was unclear for respectively 64 (67%) and 18 (64%) of the models published, resulting in unclear reporting of statistical power in 82 of the 124 models. In half of the studies which clearly described the effective sample size and number of candidate predictors, model development was underpowered.

Table 7a. Effective sample size of the included studies (reflecting statistical power)[a,b]. Numbers are column percentages and absolute numbers in parenthesis.

	Prediction as primary aim (N models 96) [b]	Prediction as secondary aim (N models 28) [b]
Considering predictors in the final model		
<5	8 (8)	0 (0)
5-10	6 (6)	25 (7)
10-15	11 (11)	4 (1)
>15	63 (60)	46 (13)
Number of patients or events not described	11 (11)	25 (7)
Considering all candidate predictors [c]		
<5	7 (7)	14 (4)
5-10	7 (7)	11 (3)
10-15	0 (0)	0 (0)
>15	19 (18)	11 (3)
Not described	67 (64)	64 (18)

[a] For continuous outcomes, the number of patients divided by the number of predictors; for dichotomous outcomes, the number of patients in the smallest category divided by the number of predictors; for time to event outcomes, the number of events divided by the number of predictors.

[b] Excluding impact and external validation studies because they require very different statistical power calculations.

[c] Number of candidate predictors was the total number of degrees of freedom (i.e. the sum of all candidate predictors, interactions, and dummy variables)

To externally validate a prediction model, a minimum effective sample size of 100 patients with and 100 without the event has been recommended[45]. Statistical power was sufficient in the majority of external validation studies (9/13 models) (Table 7b). Further, only 12/71 studies gave an explicit sample size calculation.

Table 7b. Sample size; number of events for external validation

Number of events	External validation, % (N models= 13 [a])
<100	31 (4)
100-200	0 (0)
200-400	23 (3)
>400	46 (6)

[a] Six studies externally validated 13 models.

Discussion

We have described the state of the art of current prediction research, and highlighted aspects which need improvement. We assessed the reporting and methods in clinical prediction studies rather than doing a quality appraisal. Our search indentified how many prediction studies of different types are published in high impact general medical journals. Among the 71 prediction studies published in five journals in 2008, the vast majority were predictor finding studies (n=51), apparently because risk factors remain unknown, followed by model development studies (n=14). External validation and model impact studies were rare (n=6). Study aim, design, participant selection, and definitions of outcome and predictors were generally well reported. Improvements are clearly needed in conduct and reporting issues: predictors and outcome assessment (blinding), handling of continuous predictors, testing of predictor interactions, and statistical power.

We found that 11 studies developed new prediction models (of which 3 included an external validation). 3 studies updated a previously developed model, 3 studies externally validated, and 3 investigated the impact of an existing prediction model. Development of new prediction models instead of updating or testing an existing model has been described as a trend in prediction research[8,9,15,24,36,46]. However, we found almost an identical number of model development studies, and studies which aimed to update or evaluate an existing model.

A number of basic items to describe studies were well described. These items have been identified as important by several well-known tools providing guidelines for reporting of clinical research[4,26–29]. Journals referred to these reporting guidelines in their "instructions for authors". However, sample size explanation was poorly reported despite being highlighted in reporting guidelines.

In addition in descriptions of participant selection, it often remained unclear whether participant were included in an unbiased way, notably with respect to refusals and whether all consecutive eligible patients were included. Flow diagrams were hardly ever reported, which may reflect the

difficulties of using these in prediction modeling studies due to the use of multiple analyses. The REMARK guidelines for prognostic tumor marker studies recommend using a REMARK profile table instead of a flow diagram[47].

Poor reporting of the rationale of sample size has also observed by others[3,20,48]. Further, for many studies statistical power could not be determined due to either inadequate reporting of effective sample size or the number of candidate predictors[20].

Our study examined prediction studies published in 6 high impact journals in 2008 and so is likely to be representative of higher quality published studies. Further, we assessed researchers' reporting of clinical prediction research, and not the appropriateness of the design and conduct of these studies. Conduct of clinical prediction research may be better than reported in the papers, since journals have restriction for the number of words for a paper. It is important to note that a methodologically weak or statistically underpowered study is still a poor quality study, whether or not it is well or poorly reported. However, if it is poorly reported then the reader will be unable to understand its relevance and reliability.

In this article, we reviewed reporting and methods in prediction studies. We developed a data extraction list, which may be an overview of these items. According to our observations, positive aspects of prediction research were the presence of relatively many prospective designs and adequate description of predictors. Improvement is needed in blinded assessment of predicted outcomes, handling of continuous predictors, the investigation predictor interactions, and reporting on statistical power. This review identified poor reporting and poor methods in many published prediction studies, which limit the reliability and applicability of the published literature.

References

(1) Altman DG, Riley RD (2005) Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol 2: 466-472.

(2) Altman DG (2007) Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, editors. Breast cancer. Translational therapeutic strategies. New York: New York Informa Healtcare.

(3) Altman DG, Lyman GH (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 52: 289-303.

(4) McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 97: 1180-1184.

(5) Rothwell PM (2008) Prognostic models. Pract Neurol 8: 242-253.

(6) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338: b375.

(7) Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

(8) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.

(9) Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338: b606.

(10) Steyerberg, E. W. (2009) Clinical prediction models; a practical approach to development, validation, and updating. New York: Springer.

(11) Grobbee D.E. and Hoes A.W. (2007) Clinical epidemiology: principles, methods, and applications for clinical research. Jones and Bartlett publishers.

(12) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(13) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(14) Grobbee DE (2004) Epidemiology in the right direction: the importance of descriptive research. Eur J Epidemiol 19: 741-744.

(15) Laupacis A, Sekar N, Stiell IG (1997) Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 277: 488-494.

(16) McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS (2000) Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA 284: 79-84.

(17) Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM (1999) Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282: 1061-1066.

(18) Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM (2005) Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 51: 1335-1341.

(19) Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8: 21.

(20) Mallett S, Royston P, Dutton S, Waters R, Altman DG (2010) Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 8: 20.

(21) Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV (2004) A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. J Clin Epidemiol 57: 1147-1152.

(22) Mackinnon A (2010) The use and reporting of multiple imputation in medical research - a review. J Intern Med 268: 586-593.

(23) Mushkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, Steyerberg EW (2008) A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol 61: 331-343.

(24) Perel P, Edwards P, Wentz R, Roberts I (2006) Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 6: 38.

(25) Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van d, V, Mol BW, Hompes PG (2009) Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update 15: 537-552.
(26) Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370: 1453-1457.
(27) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 138: 40-44.
(28) Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 3: 25.
(29) Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 340: c869.
(30) Hayden JA, Cote P, Bombardier C (2006) Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 144: 427-437.
(31) Hayden JA, Cote P, Steenstra IA, Bombardier C (2008) Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. J Clin Epidemiol 61: 552-560.
(32) Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JD (2001) Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Stat Med 20: 3847-3859.
(33) Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54: 774-781.
(34) Stiell IG, Wells GA (1999) Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med 33: 437-447.
(35) Toll DB, Janssen KJ, Vergouwe Y, Moons KG (2008) Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 61: 1085-1094.
(36) Reilly BM, Evans AT (2006) Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 144: 201-209.
(37) Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG (2008) Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol 8: 48.
(38) Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25: 127-141.
(39) Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR (2010) The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 340: c365.
(40) Tomlinson G, Detsky AS (2010) Composite end points in randomized trials: there is no free lunch. JAMA 303: 267-268.
(41) Lim E, Brown A, Helmy A, Mussa S, Altman DG (2008) Composite outcomes in cardiovascular research: a survey of randomized trials. Ann Intern Med 149: 612-617.
(42) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379.
(43) Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48: 1503-1510.
(44) Vittinghoff E, McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 165: 710-718.
(45) Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD (2005) Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 58: 475-483.
(46) Wasson JH, Sox HC, Neff RK, Goldman L (1985) Clinical prediction rules. Applications and methodological standards. N Engl J Med 313: 793-799.
(47) Mallett S, Timmer A, Sauerbrei W, Altman DG (2010) Reporting of prognostic studies of tumour markers: a review of published articles in relation to REMARK guidelines. Br J Cancer 102: 173-180.
(48) Bachmann LM, Puhan MA, ter RG, Bossuyt PM (2006) Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 332: 1127-1129.

Appendix: references of included studies

(1) Young Infants Clinical Signs Study Group (2008) Clinical signs that predict severe illness in children under age 2 months: a multicentre study. Lancet 371: 135-142.

(2) Acharya CR, Hsu DS, Anders CK, Anguiano A, Salter KH, Walters KS, Redman RC, Tuchman SA, Moylan CA, Mukherjee S, Barry WT, Dressman HK, Ginsburg GS, Marcom KP, Garman KS, Lyman GH, Nevins JR, Potti A (2008) Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA 299: 1574-1587.

(3) Adabag AS, Therneau TM, Gersh BJ, Weston SA, Roger VL (2008) Sudden death after myocardial infarction. JAMA 300: 2022-2029.

(4) Amin R, Widmer B, Prevost AT, Schwarze P, Cooper J, Edge J, Marcovecchio L, Neil A, Dalton RN, Dunger DB (2008) Risk of microalbuminuria and progression to macroalbuminuria in a cohort with childhood onset type 1 diabetes: prospective observational study. BMJ 336: 697-701.

(5) Bhattacharyya T, Nicholls SJ, Topol EJ, Zhang R, Yang X, Schmitt D, Fu X, Shao M, Brennan DM, Ellis SG, Brennan ML, Allayee H, Lusis AJ, Hazen SL (2008) Relationship of paraoxonase 1 (PON1) gene polymorphisms and functional activity with systemic oxidative stress and cardiovascular risk. JAMA 299: 1265-1276.

(6) Chan JA, Meyerhardt JA, Niedzwiecki D, Hollis D, Saltz LB, Mayer RJ, Thomas J, Schaefer P, Whittom R, Hantel A, Goldberg RM, Warren RS, Bertagnolli M, Fuchs CS (2008) Association of family history with cancer recurrence and survival among patients with stage III colon cancer. JAMA 299: 2515-2523.

(7) Chan PS, Krumholz HM, Nichol G, Nallamothu BK (2008) Delayed time to defibrillation after in-hospital cardiac arrest. N Engl J Med 358: 9-17.

(8) Cheyne H, Hundley V, Dowding D, Bland JM, McNamee P, Greer I, Styles M, Barnett CA, Scotland G, Niven C (2008) Effects of algorithm for diagnosis of active labour: cluster randomised trial. BMJ 337: a2396.

(9) Dehghan A, Kottgen A, Yang Q, Hwang SJ, Kao WL, Rivadeneira F, Boerwinkle E, Levy D, Hofman A, Astor BC, Benjamin EJ, van Duijn CM, Witteman JC, Coresh J, Fox CS (2008) Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. Lancet 372: 1953-1961.

(10) Diep BA, Chambers HF, Graber CJ, Szumowski JD, Miller LG, Han LL, Chen JH, Lin F, Lin J, Phan TH, Carleton HA, McDougal LK, Tenover FC, Cohen DE, Mayer KH, Sensabaugh GF, Perdreau-Remington F (2008) Emergence of multidrug-resistant, community-associated, methicillin-resistant Staphylococcus aureus clone USA300 in men who have sex with men. Ann Intern Med 148: 249-257.

(11) Fleming J, Brayne C (2008) Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. BMJ 337: a2227.

(12) Frank PI, Morris JA, Hazell ML, Linehan MF, Frank TL (2008) Long term prognosis in preschool children with wheeze: longitudinal postal questionnaire study 1993-2004. BMJ 336: 1423-1426.

(13) Freiberg JJ, Tybjaerg-Hansen A, Jensen JS, Nordestgaard BG (2008) Nonfasting triglycerides and risk of ischemic stroke in the general population. JAMA 300: 2142-2152.

(14) Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM (2008) Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. Lancet 371: 923-931.

(15) Gunnell D, Hawton K, Ho D, Evans J, O'Connor S, Potokar J, Donovan J, Kapur N (2008) Hospital admissions for self harm after discharge from psychiatric inpatient care: cohort study. BMJ 337: a2278.

(16) Gutierrez OM, Mannstadt M, Isakova T, Rauh-Hain JA, Tamez H, Shah A, Smith K, Lee H, Thadhani R, Juppner H, Wolf M (2008) Fibroblast growth factor 23 and mortality among patients undergoing hemodialysis. N Engl J Med 359: 584-592.

(17) Hall AJ, Logan JE, Toblin RL, Kaplan JA, Kraner JC, Bixler D, Crosby AE, Paulozzi LJ (2008) Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA 300: 2613-2620.

(18) Head J, Ferrie JE, Alexanderson K, Westerlund H, Vahtera J, Kivimaki M (2008) Diagnosis-specific sickness absence as a predictor of mortality: the Whitehall II prospective cohort study. BMJ 337: a1469.

(19) Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, York J, Das A, McAuley JH (2008) Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. BMJ 337: a171.

(20) Hernandez AF, Shea AM, Milano CA, Rogers JG, Hammill BG, O'Connor CM, Schulman KA, Peterson ED, Curtis LH (2008) Long-term outcomes and costs of ventricular assist devices among Medicare beneficiaries. JAMA 300: 2398-2406.

(21) Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P (2008) Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 336: 1475-1482.
(22) Holtzer R, Verghese J, Wang C, Hall CB, Lipton RB (2008) Within-person across-neuropsychological test variability and incident dementia. JAMA 300: 823-830.
(23) Imperiale TF, Glowinski EA, Lin-Cooper C, Larkin GN, Rogge JD, Ransohoff DF (2008) Five-year risk of colorectal neoplasia after negative screening colonoscopy. N Engl J Med 359: 1218-1224.
(24) Kahn SR, Shrier I, Julian JA, Ducruet T, Arsenault L, Miron MJ, Roussin A, Desmarais S, Joyal F, Kassis J, Solymoss S, Desjardins L, Lamping DL, Johri M, Ginsberg JS (2008) Determinants and time course of the postthrombotic syndrome after acute deep venous thrombosis. Ann Intern Med 149: 698-707.
(25) Kaklamani VG, Wisinski KB, Sadim M, Gulden C, Do A, Offit K, Baron JA, Ahsan H, Mantzoros C, Pasche B (2008) Variants of the adiponectin (ADIPOQ) and adiponectin receptor 1 (ADIPOR1) genes and colorectal cancer risk. JAMA 300: 1523-1531.
(26) Kerr EA, Zikmund-Fisher BJ, Klamerus ML, Subramanian U, Hogan MM, Hofer TP (2008) The role of clinical uncertainty in treatment decisions for diabetic patients with uncontrolled blood pressure. Ann Intern Med 148: 717-727.
(27) Kruijshaar ME, Watson JM, Drobniewski F, Anderson C, Brown TJ, Magee JG, Smith EG, Story A, Abubakar I (2008) Increasing antituberculosis drug resistance in the United Kingdom: analysis of National Surveillance Data. BMJ 336: 1231-1234.
(28) Kuller LH, Tracy R, Belloso W, De WS, Drummond F, Lane HC, Ledergerber B, Lundgren J, Neuhaus J, Nixon D, Paton NI, Neaton JD (2008) Inflammatory and coagulation biomarkers and mortality in patients with HIV infection. PLoS Med 5: e203.
(29) Laiyemo AO, Murphy G, Albert PS, Sansbury LB, Wang Z, Cross AJ, Marcus PM, Caan B, Marshall JR, Lance P, Paskett ED, Weissfeld J, Slattery ML, Burt R, Iber F, Shike M, Kikendall JW, Lanza E, Schatzkin A (2008) Postpolypectomy colonoscopy surveillance guidelines: predictive accuracy for advanced adenoma at 4 years. Ann Intern Med 148: 419-426.
(30) Lederle FA, Larson JC, Margolis KL, Allison MA, Freiberg MS, Cochrane BB, Graettinger WF, Curb JD (2008) Abdominal aortic aneurysm events in the women's health initiative: cohort study. BMJ 337: a1724.
(31) Limaye AP, Kirby KA, Rubenfeld GD, Leisenring WM, Bulger EM, Neff MJ, Gibran NS, Huang ML, Santo Hayes TK, Corey L, Boeckh M (2008) Cytomegalovirus reactivation in critically ill immunocompetent patients. JAMA 300: 413-422.
(32) Lin GA, Dudley RA, Lucas FL, Malenka DJ, Vittinghoff E, Redberg RF (2008) Frequency of stress testing to document ischemia prior to elective percutaneous coronary intervention. JAMA 300: 1765-1773.
(33) Loeb M, Hanna S, Nicolle L, Eyles J, Elliott S, Rathbone M, Drebot M, Neupane B, Fearon M, Mahony J (2008) Prognosis after West Nile virus infection. Ann Intern Med 149: 232-241.
(34) Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359: 2220-2232.
(35) Marcucci G, Radmacher MD, Maharry K, Mrozek K, Ruppert AS, Paschka P, Vukosavljevic T, Whitman SP, Baldus CD, Langer C, Liu CG, Carroll AJ, Powell BL, Garzon R, Croce CM, Kolitz JE, Caligiuri MA, Larson RA, Bloomfield CD (2008) MicroRNA expression in cytogenetically normal acute myeloid leukemia. N Engl J Med 358: 1919-1928.
(36) McQueen MJ, Hawken S, Wang X, Ounpuu S, Sniderman A, Probstfield J, Steyn K, Sanderson JE, Hasani M, Volkova E, Kazmi K, Yusuf S (2008) Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study. Lancet 372: 224-233.
(37. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB, Sr., Cupples LA (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359: 2208-2219.
(38) Meltzer ME, Lisman T, Doggen CJ, de Groot PG, Rosendaal FR (2008) Synergistic effects of hypofibrinolysis and genetic and acquired risk factors on the risk of a first venous thrombosis. PLoS Med 5: e97.
(39) Mermin J, Musinguzi J, Opio A, Kirungi W, Ekwaru JP, Hladik W, Kaharuza F, Downing R, Bunnell R (2008) Risk factors for recent HIV infection in Uganda. JAMA 300: 540-549.
(40) Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R, Urbauer D, Pennacchio LA, Cheng JF, Nick AM, Deavers MT, Mourad-Zeidan A, Wang H, Mueller P, Lenburg ME, Gray JW, Mok S, Birrer MJ, Lopez-Berestein G, Coleman RL, Bar-Eli M, Sood AK (2008) Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med 359: 2641-2650.

R1 (41) Montalvo G, Avanzini F, Anselmi M, Prandi R, Ibarra S, Marquez M, Armani D, Moreira JM, Caicedo C, Roncaglioni MC, Colombo F, Camisasca P, Milani V, Quimi S, Gonzabay F, Tognoni G (2008) Diagnostic evaluation of people with hypertension in low income country: cohort study of "essential" method of risk stratification. BMJ 337: a1387.
R2
R3 (42) Moylan CA, Brady CW, Johnson JL, Smith AD, Tuttle-Newhall JE, Muir AJ (2008) Disparities in liver transplantation before and after introduction of the MELD score. JAMA 300: 2371-2378.
R4
R5 (43) Nader PR, Bradley RH, Houts RM, McRitchie SL, O'Brien M (2008) Moderate-to-vigorous physical activity from ages 9 to 15 years. JAMA 300: 295-305.
R6
R7 (44) Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB, Sr., Kannel WB, Vasan RS (2008) A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med 148: 102-110.
R8
R9 (45) Peacock WF, De MT, Fonarow GC, Diercks D, Wynne J, Apple FS, Wu AH (2008) Cardiac troponin and outcome in acute heart failure. N Engl J Med 358: 2117-2126.
R10 (46) Pearce A, Law C, Elliman D, Cole TJ, Bedford H (2008) Factors associated with uptake of measles, mumps, and rubella vaccine (MMR) and use of single antigen vaccines in a contemporary UK cohort: prospective cohort study. BMJ 336: 754-757.
R11
R12 (47) Peberdy MA, Ornato JP, Larkin GL, Braithwaite RS, Kashner TM, Carey SM, Meaney PA, Cen L, Nadkarni VM, Praestgaard AH, Berg RA (2008) Survival from in-hospital cardiac arrest during nights and weekends. JAMA 299: 785-792.
R13
R14 (48) Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Poccock S, Roberts I, Shakur H, Steyerberg E, Yutthakasemsunt S (2008) Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ 336: 425-429.
R15
R16 (49) Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, Overvad K, van der Schouw YT, Spencer E, Moons KG, Tjonneland A, Halkjaer J, Jensen MK, Stegger J, Clavel-Chapelon F, Boutron-Ruault MC, Chajes V, Linseisen J, Kaaks R, Trichopoulou A, Trichopoulos D, Bamia C, Sieri S, Palli D, Tumino R, Vineis P, Panico S, Peeters PH, May AM, Bueno-de-Mesquita HB, van Duijnhoven FJ, Hallmans G, Weinehall L, Manjer J, Hedblad B, Lund E, Agudo A, Arriola L, Barricarte A, Navarro C, Martinez C, Quiros JR, Key T, Bingham S, Khaw KT, Boffetta P, Jenab M, Ferrari P, Riboli E (2008) General and abdominal adiposity and risk of death in Europe. N Engl J Med 359: 2105-2120.
R17
R18
R19
R20
R21 (50) Rawstron AC, Bennett FL, O'Connor SJ, Kwok M, Fenton JA, Plummer M, de TR, Owen RG, Richards SJ, Jack AS, Hillmen P (2008) Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. N Engl J Med 359: 575-583.
R22
R23 (51) Righini M, Le GG, Aujesky D, Roy PM, Sanchez O, Verschuren F, Rutschmann O, Nonent M, Cornuz J, Thys F, Le Manach CP, Revel MP, Poletti PA, Meyer G, Mottier D, Perneger T, Bounameaux H, Perrier A (2008) Diagnosis of pulmonary embolism by multidetector CT alone or combined with venous ultrasonography of the leg: a randomised non-inferiority trial. Lancet 371: 1343-1352.
R24
R25
R26 (52) Ro KE, Gude T, Tyssen R, Aasland OG (2008) Counselling for burnout in Norwegian doctors: one year cohort study. BMJ 337: a2004.
R27 (53) Sasson C, Hegg AJ, Macy M, Park A, Kellermann A, McNally B (2008) Prehospital termination of resuscitation in cases of refractory out-of-hospital cardiac arrest. JAMA 300: 1432-1438.
R28
R29 (54) Sattar N, McConnachie A, Shaper AG, Blauw GJ, Buckley BM, de Craen AJ, Ford I, Forouhi NG, Freeman DJ, Jukema JW, Lennon L, Macfarlane PW, Murphy MB, Packard CJ, Stott DJ, Westendorp RG, Whincup PH, Shepherd J, Wannamethee SG (2008) Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. Lancet 371: 1927-1935.
R30
R31 (55) Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N, Yuen ST, Chan TL, Kwong DL, Au GK, Liu CG, Calin GA, Croce CM, Harris CC (2008) MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. JAMA 299: 425-436.
R32
R33 (56) Schlenk RF, Dohner K, Krauter J, Frohling S, Corbacioglu A, Bullinger L, Habdank M, Spath D, Morgan M, Benner A, Schlegelberger B, Heil G, Ganser A, Dohner H (2008) Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. N Engl J Med 358: 1909-1918.
R34
R35 (57) Sekhri N, Feder GS, Junghans C, Eldridge S, Umaipalan A, Madhu R, Hemingway H, Timmis AD (2008) Incremental prognostic value of the exercise electrocardiogram in the initial assessment of patients with suspected angina: cohort study. BMJ 337: a2240.
R36
R37
R38 (58) Smith GC, Celik E, To M, Khouri O, Nicolaides KH (2008) Cervical length at mid-pregnancy and the risk of primary cesarean delivery. N Engl J Med 358: 1346-1353.
R39

(59) Stern DA, Morgan WJ, Halonen M, Wright AL, Martinez FD (2008) Wheezing and bronchial hyper-responsiveness in early childhood as predictors of newly diagnosed asthma in early adulthood: a longitudinal birth-cohort study. Lancet 372: 1058-1064.
(60) Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI (2008) Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 5: e165.
(61) Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med 148: 337-347.
(62) Tyson JE, Parikh NA, Langer J, Green C, Higgins RD (2008) Intensive care for extreme prematurity--moving beyond gestational age. N Engl J Med 358: 1672-1681.
(63) Tzemos N, Therrien J, Yip J, Thanassoulis G, Tremblay S, Jamorski MT, Webb GD, Siu SC (2008) Outcomes in adults with bicuspid aortic valves. JAMA 300: 1317-1325.
(64) van Veen M, Steyerberg EW, Ruige M, van Meurs AH, Roukema J, van der LJ, Moll HA (2008) Manchester triage system in paediatric emergency care: prospective observational study. BMJ 337: a1501.
(65) Vestergaard M, Pedersen MG, Ostergaard JR, Pedersen CB, Olsen J, Christensen J (2008) Death in children with febrile seizures: a population-based cohort study. Lancet 372: 457-463.
(66) Vidula H, Tian L, Liu K, Criqui MH, Ferrucci L, Pearce WH, Greenland P, Green D, Tan J, Garside DB, Guralnik J, Ridker PM, Rifai N, McDermott MM (2008) Biomarkers of inflammation and thrombosis as predictors of near-term mortality in patients with peripheral arterial disease: a cohort study. Ann Intern Med 148: 85-93.
(67) Viros A, Fridlyand J, Bauer J, Lasithiotakis K, Garbe C, Pinkel D, Bastian BC (2008) Improving melanoma classification by integrating genetic and morphologic features. PLoS Med 5: e120.
(68) Wang NC, Maggioni AP, Konstam MA, Zannad F, Krasa HB, Burnett JC, Jr., Grinfeld L, Swedberg K, Udelson JE, Cook T, Traver B, Zimmer C, Orlandi C, Gheorghiade M (2008) Clinical implications of QRS duration in patients hospitalized with worsening heart failure and reduced left ventricular ejection fraction. JAMA 299: 2656-2666.
(69) Xie J, Brayne C, Matthews FE (2008) Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up. BMJ 336: 258-262.
(70) Zethelius B, Berglund L, Sundstrom J, Ingelsson E, Basu S, Larsson A, Venge P, Arnlov J (2008) Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. N Engl J Med 358: 2107-2116.
(71) Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Gronberg H (2008) Cumulative association of five genetic variants with prostate cancer. N Engl J Med 358: 910-919

Reporting of statistical methods, predictive performance and validation techniques in clinical prediction research: a systematic review

NPA Zuithoff*[1], W Bouwmeester*[1], S Mallett[2,3], MI Geerlings[1], EW Steyerberg[4], Y Vergouwe[1], DG Altman[3], KGM Moons[1**].

*equally contributed

Affiliations

[1] Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, the Netherlands

[2] Department of Primary Care, University of Oxford, Oxford, UK

[3] Centre for Statistics in Medicine, University of Oxford, Wolfson College Annexe, Oxford, UK

[4] Department of Public Health, Erasmus MC, Rotterdam, the Netherlands

Provisionally accepted in adapted form by PLoS Medicin

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

In recent years, there has been an increasing interest in the methodology of prediction research[1–16]. Prediction research includes both diagnostic prediction studies studying the ability of diagnostic variables or test results to predict the presence or absence of a certain diagnosis, and prognostic studies studying predictors of the future occurrence of outcomes[6,11,15]. Both types of prediction research may investigate the predictive effect of a single variable (or predictor or test), seek to identify the independently contributing predictors among multiple candidate predictors, intend to develop a new prediction model, to validate the predictive accuracy of an existing model or to quantify its impact on clinical decision making and patient outcome[6]. As the use of poor methods leads to biased results, many authors provided methodological recommendations and stressed the importance of pre-defining key aspects and methods of a prediction studies including a clear aim, proper study population and subject selection, clinically relevant outcomes, proper selection of candidate and final predictors, required sample size and a proper statistical analysis[2,7,10,12,13,15,17–23]. Few authors addressed the poor reporting of prediction modelling studies[15,24–33]. Reporting guidelines for prediction research have received limited attention, with just a few exceptions[34,35]. We therefore performed a comprehensive literature review of all articles published in high general medical journals to assess whether prediction research in the recent literature was conducted and reported according to methodological recommendations. We focused on all types of clinical prediction studies and all methodological issues that are considered to be important in prediction research, rather than on specific types of outcomes (such as dichotomous outcomes[27]), specific methodological issues (such as missing data[36]), or specific disease areas (e.g. oncology)[24–26,28,37]. In the preceding chapter (see chapter 6) we focus on the reporting of the study aims, design, selection of participants, definition and measurement of the outcomes and predictors, and statistical power. In this second chapter we focus on the reporting of the statistical methods, including e.g. the selection of predictors and handling of missing values, whether model validation was carried out, and the reporting of the results including of predictive performance measures.

Methods

We hand searched the 6 highest impact general medicine journals - The New England Journal of Medicine, Lancet, Journal of the American Medical Association, Annals of Internal Medicine, PLoS Medicine, and British Medical Journal - published in 2008. Details of the literature search, inclusion criteria and the development of the item list are reported in the preceding chapter (see chapter 6). For this second paper, we extracted items regarding the selection of predictors both prior to the analysis (candidate predictors) and within the statistical analysis (independent predictors), criteria used for predictor selection, the reporting and handling of missing values, the presentation of the results including of predictive performance measures, and whether any

validity assessments were reported (Box 1). Items were assessed as present, absent, not applicable or unclear. In the results presented here, items coded unclear were counted as absent.

Reliability

Two authors (WB, NPAZ) independently extracted the data of the included studies. In case of doubt, items were discussed with the other authors (KGMM, YV). The validity of the extraction was assessed by calculating the percentage of overall agreement between the two reviewers.

Analysis

For this paper, we grouped results by type of clinical prediction research, and whether the prediction analysis was a primary or secondary aim of the study. We distinguished 5 types of clinical prediction research: (1) Predictor finding studies[3,6,38], studies that aim to discover or explore which predictors or variables are independently associated with an outcome; (2) Model development studies without applying a validation in other individuals than used for the model development, (so-called external model validation): (3) Model development studies with external model validation; (4) Pure external model validation studies, with or without model updating or adjustment and (5) studies aimed at quantifying the impact of prediction models on decision making or patient outcomes. An exact description of these types is given in the previous chapter(see chapter 6).

We focused on types 1 to 4 studies only. Studies that exclusively report on the impact of prediction models were excluded, as these studies generally require a comparative design and thus very different methodological issues apply to these type of studies as compared to the other 4 types.

Results

We retrieved 71 papers for full text review which included 51 papers on predictor finding studies and 17 papers on the development or validation of one or more prediction models. Only 6 studies addressed an external validation (n=3) or impact (n=3) of a previously developed model. Prediction was the primary aim in 48 of the 71 papers and the secondary aim in 17. The included studies are listed in the appendix.

The median agreement between the two reviewers was 92% (IQR 75%-100%). Lower levels of agreement for some items were due to ambiguities in the reviewed articles. Differences were resolved after discussion with the third independent reviewer.

Selection of predictors

Adequate reporting of the predictor selection is important, as the methods, the number of candidate predictors and the choice of specific predictors, can all influence the predictor selection in the final multivariable analyses and thus the interpretation of the results. This issue is

not specific to prediction studies but also arises in causal research, although here variables to be included in the multivariable modelling are usually referred to as confounders. Ideally, candidate predictors are selected based on theoretical or clinical understanding. No specific method is generally recommended for predictor selection, neither for selecting the candidate predictors nor for selecting the independent predictors from this set of candidate predictors in the statistical analyses. Some methods, such as the selection based on univariable statistics, increase the chance on biased results and are therefore not recommended[12,13,39,40].

In multivariable analyses, predictors are often selected, to conclude on the independent predictors or for inclusion in a final prediction model, based on backward or forward selection using, for instance a significance level of 0.05. Various authors warned for the increase in chance of over-optimistic results[10,12,13], especially when there are relatively few outcome events and many predictors analysed[12,13]. However, there is no clear best alternative.

The selection of which candidate predictors were considered in the study was described in 36 (75%) of the studies where prediction was the primary aim, and in 8 (47%) of the 'secondary aim prediction studies' (Table 1). The majority of studies with prediction as primary aim (n=34, 71%) selected candidate predictors based on existing literature, whereas this was less often done (n=5, 29%) in studies with prediction as secondary aim. A pre-selection based on univariable analysis was used in 6 (13%) primary aim studies and in 4 (24%) secondary aim studies (Table 1).

The method of selection of predictors within the multivariable statistical model was not described in 13 (27%) primary aim studies and in 6 (35%) secondary aim studies (Table 1). Backward selection was reported in 17% of primary aim studies and in 18% of the secondary aim studies, whereas forward selection was reported in 6% and 0%, respectively. The most commonly reported criterion for predictor selection in the multivariable was the p-value <0.05 in 21% of all studies (29% of studies that applied selection). Other criteria, such as Akaike's Information Criterion or R^2, were used less often (Table 1). Twelve studies (18% of all studies) investigated the added value of specific predictors by overriding automatic variable selection to include known predictors in the model. Twentyseven (42%) studies also included known predictors regardless of statistical significance in the particular patient dataset.

Missing values

Missing values rarely occur completely at random. Commonly missing values are related to observed patient or disease characteristics. Exclusion of patients with missing values will therefore not only lead to loss of statistical power, but often also to biased results[10,12,13,20,21,41,42]. Imputation, notably multiple imputation, of missing values is often advocated to preserve power and obtain less biased results, on the assumption that the reason for the missing data is not entirely due to non-observed information (i.e. data are not "missing not at random"). When there are few missing observations, for example < 5% of the individual values in the data, sometimes simple methods are advocated such as single imputation or imputation of the mean[13,20,43].

Table 1. Method of predictor selection, stratified by whether prediction was the primary or secondary study aim [a]. Numbers are column percentages and absolute numbers between parenthesis.

	Prediction as primary aim (N=48)	Prediction as secondary aim (N=17)	Total (N=65)
Selection of predictors for inclusion in the multivariable analysis			
Not based on statistical analysis [b]			
Method described [c]	75 (36)	47 (8)	68 (44)
Literature based	71 (34)	29 (5)	60 (39)
A priori hypothesis/clinical reasoning	29 (14)	29 (5)	29 (19)
Based on statistical analysis			
Screening by univariable analysis	13 (6)	24 (4)	15 (10)
Method of predictor selection used within multivariable analysis [b]			
Backward selection	17 (8)	18 (3)	17 (11)
Forward selection	6 (3)	0 (0)	5 (3)
Added value of a specific predictor to existing predictors or model [d]	25 (12)	0 (0)	18 (12)
All predictors included regardless of statistical significance	40 (19)	47 (8)	42 (27)
Similar predictors combined [e]	17 (8)	6 (1)	11 (7)
Method not described	27 (13)	35 (6)	29 (19)
Selection of predictors in multivariable analyses based on [f]			
p-values <0.05 or smaller	21; 29 (10)	12; 18 (2)	18; 26 (12)
p-values higher than 0.05	4; 6 (2)	12; 18 (2)	6; 9 (4)
Akaike's Information Criterion	4; 6 (2)	0; 0 (0)	3; 4 (2)
Bayesian Information Criterion	2; 6 (1)	6; 9 (1)	3; 4 (2)
Explained variance (R^2)	4; 6 (2)	0; 0 (0)	3; 4 (2)
Change in c-statistic	10; 14 (5)	0; 0 (0)	9; 13 (6)

[a] External validation studies (N=3) were excluded from this table as these issues are not applicable for these type of studies. Hence, N=68.
[b] More than one method may be used within a study, percentages do not add up to 100%.
[c] Percentage (number) of studies that reported the applied method for selecting which predictors were included in the multivariable analyses, if it was not based on statistical analysis (i.e. univariable predictor-outcome associations).
[d] Predictor inclusion in multivariable model was prespecified as the specific aim was to quantify its added predictive value to existing predictors.
[e] E.g. systolic and diasystolic blood pressure combined to mean blood pressure.
[f] For the items below, two percentages are given. The first is the percentage of all studies (48, 17 and 65 respectively), the second is the percentage of all studies that applied some type of predictor selection in the multivariable analysis (35, 11 and 46 respectively; the excluded studies did not apply any predictor selection in the multivariable analysis but pre-specified the final model).

Missing values were mostly reported per predictor, in 56% of all studies (Table 2). The number of patients with any missing values was reported less often, in 22% of all studies. Lost to follow up was reported in 42% of the studies where this was applicable. An analysis of participants with complete data (i.e. complete case analysis), ignoring all patients with one or more missing values, was performed in 89%. By comparison, multiple imputation, the most rigorous strategy for dealing with missing values, was used in 8% of all studies. In another strategy, the missing indicator method, a new dummy or indicator (0/1) variable is created for every predictor with missing values, with '1' indicating a missing on the original predictor and '0' indicating an observed value. This predictor is included in the multivariable analysis. Even though this method is known to lead to biased results in almost all observational studies[13,20,44,45], it was still used in 14% of all studies.

Presentation of results

Most guidelines, such as the STROBE guidelines for the reporting of observational studies or the REMARK guidelines for tumour marker prognostic studies, specifically advise to report both unadjusted (i.e. from univariable analysis, yielding the association of each candidate predictor with the outcome) and adjusted (i.e. from a multivariable analysis) results[4,35,46]. Presenting results from both analyses allows readers insight in the predictor selection strategies and to determine the influence of the adjustment for other predictors. For prediction studies that have applied predictor selection methods in the multivariable analyses, i.e. the presentation of a 'full' model, a model that includes all predictors considered, is also valuable, as it gives readers insight into the multivariable selection process.

Few studies reported adjusted or unadjusted results of the full model with all predictors considered (Table 3). The majority of clinical prediction studies, 65% of the predictor finding studies and 79% of the development studies, reported regression model coefficients or effect estimates, such as odds ratios, hazard ratios, of the model after predictor selection (the final model). The remaining studies reported only the effects of a specific predictor of interest and omitted the results for the other predictors included in the analysis.

Table 2. Handling of missing values, stratified by whether prediction was the primary or secondary study aim. Numbers are column percentages and absolute numbers between parenthesis.

	Prediction as primary aim (N=48) and external validation studies (N=3) combined	Prediction as secondary aim (N=17)	Total (N=68)
Reporting of missing data [b]			
Not reported or unclear	35 (18)	53 (9)	40 (27)
Number of participants with missing values	25 (13)	12 (2)	22 (15)
Number of missing values per predictor	59 (30)	47 (8)	56 (38)
Number of loss-to-follow-up [c]	40 (17)	50 (7)	42 (24)
Methods used for handling of missing data [d]			
Complete case analysis [e]	90 (44)	88 (15)	89 (59)
Predictor with missing values omitted	2 (1)	12 (2)	5 (3)
Missing indicator method	14 (7)	12 (2)	14 (9)
Single imputation	2 (1)	6 (1)	3 (2)
Multiple imputation	10 (5)	0 (0)	8 (5)
Sensitivity analysis [f]	6 (3)	24 (4)	11 (7)
Not reported or unclear	49 (24)	65 (11)	53 (35)

[a] Studies that report on external validation only (n=3) were included in this table under Prediction as primary aim.
[b] Some studies reported more than one item. Hence, the summation of the percentages is over 100%.
[c] Cross-sectional studies were excluded for this item (item not applicable) including 8 studies with prediction as primary aim, 3 as secondary aim.
[d] More than one method could be applied, and complete case analysis was assumed for 15 studies that unclearly described handling of missing values. Hence, the summation of the percentages is over 100%. Items were not applicable for 2 primary aim studies which had no missing values.
[e] Studies which analysed only the participants with completely observed data.
[f] Investigators assumed for instance that among participants who did not undergo follow-up colonoscopy, the detection rates for any adenoma and for an advanced adenoma ranged from half to twice the rates among participants who underwent follow-up colonoscopy, reference [23] in the appendix.

Table 3. Presentation of the results, stratified by type of prediction study. Numbers are column percentages and absolute numbers between parenthesis.[a]

	Predictor finding studies (N=51)	Development studies (N=14)	Total (N=65)
Unadjusted (univariable) association of each candidate predictor with the outcome	18 (9)	21 (3)	18 (12)
Unadjusted association only of the predictors eventually included in the final model(i.e. after predictor selection)	37 (19)	29 (4)	35 (23)
Adjusted associations of each predictor in the full or initial multivariable model	18 (9)	29 (4)	20 (13)
Adjusted associations of each predictor in the final multivariable model	65 (33)	79 (11)	68 (44)
Simplified risk score / nomogram / score chart	4 (2)	36 (5)	11 (7)

[a] Impact and external validation studies (N=6) were excluded from this table as these items were not applicable. Hence, total N=65.

Model performance and internal and external validity

The assessment of the predictive performance of a model is important to understand how predictions from the model perform, in comparison to observed results. Predictive performance of a model can be assessed on the same data that was used to generate the results (referred to as the apparent performance in the development data set), or in random subsamples of the development data set, referred to as internal validation of the prediction model[2,8,10,29,47,48]. Validating this model's predictive performance in new subject data (i.e. other subjects than used for the model development or internal validation), is the most rigorous form of model validity assessment and is referred to as external validation[8,10,47,49].

In prediction research, two types of prediction performance measures are distinguished: calibration, which is the agreement between predicted outcome and observed outcome, and discrimination, which is the ability to separate patients with and without the outcome of interest[13,50]. Discrimination measures are applicable for dichotomous and time-to-event outcomes. In addition, overall measures for discrimination and calibration (e.g. the R^2 and Brier scores) may also be reported.

Calibration was reported in few studies. If done, the Hosmer-Lemeshow statistic was the most often reported calibration measure (9%, n=6). Discrimination was assessed with the c-statistic or area under the ROC curve in 12% of the predictor finding studies and in 80% of the model development and external validation studies. R^2 and Brier score were reported in very few studies. Internal validation was performed in 33% (n=5) of the 14 model development studies, and external validation in only 4.

Table 4. Model performance measures, stratified by type of prediction study. Numbers are column percentages and absolute numbers between parenthesis.[a]

	Predictor finding studies (N=51)	Development (N=14) and external validation (N=1) [b] studies combined	Total (N=66) [b]
Calibration measures			
Calibration plot	0 (0)	27 (4)	6 (4)
Calibration intercept and slope	0 (0)	0 (0)	0 (0)
Hosmer-Lemeshow statistic	4 (2)	27 (4)	9 (6)
Discrimination measures			
C-statistic/AUC-ROC	12 (6)	80 (12)	27 (18)
Classification			
NRI	2 (1)	40 (6)	11 (7)
Sensitivity/specificity	2 (1)	7 (1)	3 (2)
Other	2 (1)	33 (5)	9 (6)
Overall performance measures			
Brier	0 (0)	7 (1)	2 (1)
R^2	8 (4)	13 (2)	9 (6)
Validity assessment			
Apparent [c]	18 (9)	60 (9)	27 (18)
Internal with jack-knife	0 (0)	7 (1)	2 (1)
Internal with (random) split sample	0 (0)	13 (2)	3 (2)
Internal with bootstrapping techniques	4 (2)	13 (2)	6 (4)
External	0 (0)	27 (4)	6 (4)

[a] The percentages can add up over 100% because development studies commonly reported >1 performance measure or validity assessment.
[b] Two studies on external validation were excluded from this table, because risk stratification tools were evaluated which did not give predicted probabilities (the Manchester triage system and predictive life support tools; reference [53,64] in the appendix). Hence, almost all items are not applicable, as these studies reported sensitivities and specificities only.
[c] The predictive performance (e.g. C-statistic, calibration or NRI) of the prediction model as estimated from the same data as from which the model was developed.
NRI= net reclassification index. AUC-ROC= area under the receiver operation characteristic curve

Discussion

We have described the reporting in current prediction research, and highlighted aspects which need improvement. In this second of two papers, we focused on the reporting and conduct of the statistical analyses and results sections.

Even though the selection of predictors was well described for many clinical prediction research studies, statistical methods used were often poor. Univariable pre-selection was used in 10

studies, and automated predictor selection in multivariable analyses based on pure p-values (notably < 0.05) was very often used. These methods may notably be problematic for studies with low numbers of events or outcomes and many candidate predictors. As the exact number of events per candidate predictor could almost never be assessed (see chapter 6) in the included studies, it was not possible to determine if results were biased[13]. Several studies, however, did not rely exclusively on automatic stepwise selection of predictors during model development, but also included established predictors in their model regardless of statistical significance in their dataset.

Almost all studies reported missing data present in their dataset. Despite many methodological recommendations about its potential for yielding biased results[10,12,13,20,21,41,42], complete case analysis was by far the most commonly used approach to handle missing values. Advocated methods, such as multiple imputation, were applied and reported in only very few studies, although this may be due to the relative recent recommendation of these methods. As the reason of the missing values were insufficiently described in most studies that applied a complete case analyses, it was impossible to judge if imputation methods would indeed have been appropriate.

Most studies correctly reported the predictor effects derived from the final multivariable analyses. However, only few studies also reported results of the univariable analyses. As noted in e.g. the REMARK guideline[4], a comprehensive reporting of both univariable and multivariable analyses would at least allow readers to evaluate the adjustment for other predictors.

We observed much variation in the reporting of the predictive performance measures. The c-statistic or area under the ROC-curve in case of dichotomous outcomes, a measure of discrimination, was the most frequently reported, whereas measures of calibration (such as calibration slope) and overall measures of fit were rarely reported. Calibration measures are important, notably in prediction model validation studies, to judge whether the predicted probabilities indeed match observed frequencies of the outcome under study, indicating a well calibrated model.

The majority of model development studies reported the predictive performance in the development data. This apparent model performance, however, is generally too optimistic, as the model has been tailored to the data set at hand. This optimisme is more likely when the number of candidate predictors starts to outweigh the number of study subjects[10,12,51]. The extent of this optimism may be estimated with so-called internal validation techniques[10,12,39,51], but use of these techniques was rare. Similarly, only very few model development studies reported an external validation of the model in the same paper. Accordingly, the generalisability of the performance of all these reported models, especially in studies where prediction was the primary aim, can hardly be evaluated by readers.

We examined prediction studies published in 6 high impact journals and so is likely to be representative of higher quality published studies. Our search comprised the year 2008. Reporting may obviously have been improved in the past two years. However, as no major reporting guidelines of this type of research have been produced in the past two years, we believe the results are still

representative for the current situation of prediction research. Hopefully, very recent publication of the GRIPS and REMARK statements, even though focussed on specific types of studies[34], may improve reporting of future prediction research in general. We also note that the conduct of the addressed prediction studies may have been better than reported in the papers, since journals have restriction for the number of words and the number of tables for a publication. However, where journals allow publication of supplementary online information there is little excuse for omitting important information needed for readers. This information, if available, was assessed in this review.

Most other reviews of prediction studies have been limited to specific outcomes (e.g. dichotomous outcomes[27]), specific methodological issues (such as missing data[36]) or a specific disease area (e.g. oncology[24,25], traumatic brain injury[26,28]). We studied a larger variety of different types of prediction research, ranging from predictor finding studies to external prediction model validation studies. Furthermore, previous reviews as well as other broad reviews[15,29,30] did not incorporate all items we have included here, or reported items in different ways. These differences limit the comparison with these existing reviews to some observations. The high reporting of selection of predictors we found compares favourably with other reviews[15,24–30]. This may suggest that reporting of prediction studies is progressing to the recommendations made in several guidelines. Also, previous reviews reported a high percentage of studies that used complete case analysis rather than applying advanced methods for missing data and they similarly observed a low percentage of studies that reported internal validation (including split sample, jack-knife or bootstrap procedures) or external validation results.

To conclude, we have identified in general poor reporting and the use of poor methods in many publications on prediction studies, which limit the reliability and generalisability of their results. We did find good reporting of the selection of predictors, both prior to the statistical analysis as well as within the analysis. However, improvement is notably needed in the reporting of the amount of missing data, the presentation of the results of multivariable analysis, and the methods used to quantify and validate the predictive performance of prediction models. Only a very small minority of the papers described an external validation of a prediction model.

References

(1) Altman DG, Riley RD (2005) Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol 2: 466-472.

(2) Altman DG (2007) Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, editors. Breast cancer. Translational therapeutic strategies. New York: New York Informa Healtcare.

(3) Altman DG, Lyman GH (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 52: 289-303.

(4) McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 97: 1180-1184.

(5) Rothwell PM (2008) Prognostic models. Pract Neurol 8: 242-253.

(6) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338: b375.

(7) Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

(8) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.

(9) Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338: b606.

(10) Steyerberg, E. W. (2009) Clinical prediction models; a practical approach to development, validation, and updating. New York: Springer.

(11) Grobbee D.E. and Hoes A.W. (2007) Clinical epidemiology: principles, methods, and applications for clinical research. Jones and Bartlett publishers.

(12) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(13) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(14) Grobbee DE (2004) Epidemiology in the right direction: the importance of descriptive research. Eur J Epidemiol 19: 741-744.

(15) Laupacis A, Sekar N, Stiell IG (1997) Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 277: 488-494.

(16) McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS (2000) Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA 284: 79-84.

(17) Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM (1999) Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282: 1061-1066.

(18) Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM (2005) Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 51: 1335-1341.

(19) Concato J, Feinstein AR, Holford TR (1993) The risk of determining risk with multivariable models. Ann Intern Med 118: 201-210.

(20) Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59: 1087-1091.

(21) Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142: 1255-1264.

(22) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379.

(23) Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48: 1503-1510.

(24) Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8: 21.

(25) Mallett S, Royston P, Dutton S, Waters R, Altman DG (2010) Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 8: 20.

(26) Mushkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, Steyerberg EW (2008) A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol 61: 331-343.
(27) Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV (2004) A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. J Clin Epidemiol 57: 1147-1152.
(28) Perel P, Edwards P, Wentz R, Roberts I (2006) Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 6: 38.
(29) Reilly BM, Evans AT (2006) Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 144: 201-209.
(30) Wasson JH, Sox HC, Neff RK, Goldman L (1985) Clinical prediction rules. Applications and methodological standards. N Engl J Med 313: 793-799.
(31) Hier DB, Edelstein G (1991) Deriving clinical prediction rules from stroke outcome research. Stroke 22: 1431-1436.
(32) Jacob M, Lewsey JD, Sharpin C, Gimson A, Rela M, van der Meulen JH (2005) Systematic review and validation of prognostic models in liver transplantation. Liver Transpl 11: 814-825.
(33) Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM (2004) Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. Ann Thorac Surg 77: 2232-2237.
(34) Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ (2011) Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Clin Invest 41: 1004-1009.
(35) Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370: 1453-1457.
(36) Mackinnon A (2010) The use and reporting of multiple imputation in medical research - a review. J Intern Med 268: 586-593.
(37) Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van d, V, Mol BW, Hompes PG (2009) Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update 15: 537-552.
(38) Hayden JA, Cote P, Steenstra IA, Bombardier C (2008) Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. J Clin Epidemiol 61: 552-560.
(39) Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54: 774-781.
(40) Sun GW, Shook TL, Kay GL (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 49: 907-916.
(41) Burton A, Altman DG (2004) Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer 91: 4-8.
(42) Gorelick MH (2006) Bias arising from missing data in predictive models. J Clin Epidemiol 59: 1115-1123.
(43) Marshall A, Altman DG, Holder RL (2010) Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 10:112.: 112.
(44) Knol MJ, Janssen KJ, Donders AR, Egberts AC, Heerdink ER, Grobbee DE, Moons KG, Geerlings MI (2010) Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 63: 728-736.
(45) White IR, Thompson SG (2005) Adjusting for partially missing baseline measurements in randomized trials. Stat Med 24: 993-1007.
(46) Vandenbroucke JP, Von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med 4: e297.
(47) Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453-473.
(48) Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med 130: 515-524.
(49) Sauerbrei W (1999) The use of resampling methods to simplify regression models in medical statistics. Applied statistics 48: 313-329.
(50) Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD (2002) Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol 20: 96-107.
(51) Van Houwelingen JC, Le Cessie S. (1990) Predictive value of statistical models. Stat Med 9: 1303-1325.

Appendix: references of included studies

(1) Young Infants Clinical Signs Study Group (2008) Clinical signs that predict severe illness in children under age 2 months: a multicentre study. Lancet 371: 135-142.

(2) Acharya CR, Hsu DS, Anders CK, Anguiano A, Salter KH, Walters KS, Redman RC, Tuchman SA, Moylan CA, Mukherjee S, Barry WT, Dressman HK, Ginsburg GS, Marcom KP, Garman KS, Lyman GH, Nevins JR, Potti A (2008) Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA 299: 1574-1587.

(3) Adabag AS, Therneau TM, Gersh BJ, Weston SA, Roger VL (2008) Sudden death after myocardial infarction. JAMA 300: 2022-2029.

(4) Amin R, Widmer B, Prevost AT, Schwarze P, Cooper J, Edge J, Marcovecchio L, Neil A, Dalton RN, Dunger DB (2008) Risk of microalbuminuria and progression to macroalbuminuria in a cohort with childhood onset type 1 diabetes: prospective observational study. BMJ 336: 697-701.

(5) Bhattacharyya T, Nicholls SJ, Topol EJ, Zhang R, Yang X, Schmitt D, Fu X, Shao M, Brennan DM, Ellis SG, Brennan ML, Allayee H, Lusis AJ, Hazen SL (2008) Relationship of paraoxonase 1 (PON1) gene polymorphisms and functional activity with systemic oxidative stress and cardiovascular risk. JAMA 299: 1265-1276.

(6) Chan JA, Meyerhardt JA, Niedzwiecki D, Hollis D, Saltz LB, Mayer RJ, Thomas J, Schaefer P, Whittom R, Hantel A, Goldberg RM, Warren RS, Bertagnolli M, Fuchs CS (2008) Association of family history with cancer recurrence and survival among patients with stage III colon cancer. JAMA 299: 2515-2523.

(7) Chan PS, Krumholz HM, Nichol G, Nallamothu BK (2008) Delayed time to defibrillation after in-hospital cardiac arrest. N Engl J Med 358: 9-17.

(8) Cheyne H, Hundley V, Dowding D, Bland JM, McNamee P, Greer I, Styles M, Barnett CA, Scotland G, Niven C (2008) Effects of algorithm for diagnosis of active labour: cluster randomised trial. BMJ 337: a2396.

(9) Dehghan A, Kottgen A, Yang Q, Hwang SJ, Kao WL, Rivadeneira F, Boerwinkle E, Levy D, Hofman A, Astor BC, Benjamin EJ, van Duijn CM, Witteman JC, Coresh J, Fox CS (2008) Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. Lancet 372: 1953-1961.

(10) Diep BA, Chambers HF, Graber CJ, Szumowski JD, Miller LG, Han LL, Chen JH, Lin F, Lin J, Phan TH, Carleton HA, McDougal LK, Tenover FC, Cohen DE, Mayer KH, Sensabaugh GF, Perdreau-Remington F (2008) Emergence of multidrug-resistant, community-associated, methicillin-resistant Staphylococcus aureus clone USA300 in men who have sex with men. Ann Intern Med 148: 249-257.

(11) Fleming J, Brayne C (2008) Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. BMJ 337: a2227.

(12) Frank PI, Morris JA, Hazell ML, Linehan MF, Frank TL (2008) Long term prognosis in preschool children with wheeze: longitudinal postal questionnaire study 1993-2004. BMJ 336: 1423-1426.

(13) Freiberg JJ, Tybjaerg-Hansen A, Jensen JS, Nordestgaard BG (2008) Nonfasting triglycerides and risk of ischemic stroke in the general population. JAMA 300: 2142-2152.

(14) Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM (2008) Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. Lancet 371: 923-931.

(15) Gunnell D, Hawton K, Ho D, Evans J, O'Connor S, Potokar J, Donovan J, Kapur N (2008) Hospital admissions for self harm after discharge from psychiatric inpatient care: cohort study. BMJ 337: a2278.

(16) Gutierrez OM, Mannstadt M, Isakova T, Rauh-Hain JA, Tamez H, Shah A, Smith K, Lee H, Thadhani R, Juppner H, Wolf M (2008) Fibroblast growth factor 23 and mortality among patients undergoing hemodialysis. N Engl J Med 359: 584-592.

(17) Hall AJ, Logan JE, Toblin RL, Kaplan JA, Kraner JC, Bixler D, Crosby AE, Paulozzi LJ (2008) Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA 300: 2613-2620.

(18) Head J, Ferrie JE, Alexanderson K, Westerlund H, Vahtera J, Kivimaki M (2008) Diagnosis-specific sickness absence as a predictor of mortality: the Whitehall II prospective cohort study. BMJ 337: a1469.

(19) Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, York J, Das A, McAuley JH (2008) Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. BMJ 337: a171.

(20) Hernandez AF, Shea AM, Milano CA, Rogers JG, Hammill BG, O'Connor CM, Schulman KA, Peterson ED, Curtis LH (2008) Long-term outcomes and costs of ventricular assist devices among Medicare beneficiaries. JAMA 300: 2398-2406.

R1 (21) Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P (2008) Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 336: 1475-1482.

R2 (22) Holtzer R, Verghese J, Wang C, Hall CB, Lipton RB (2008) Within-person across-neuropsychological test variability and incident dementia. JAMA 300: 823-830.

R3 (23) Imperiale TF, Glowinski EA, Lin-Cooper C, Larkin GN, Rogge JD, Ransohoff DF (2008) Five-year risk of colorectal neoplasia after negative screening colonoscopy. N Engl J Med 359: 1218-1224.

R4 (24) Kahn SR, Shrier I, Julian JA, Ducruet T, Arsenault L, Miron MJ, Roussin A, Desmarais S, Joyal F, Kassis J, Solymoss S, Desjardins L, Lamping DL, Johri M, Ginsberg JS (2008) Determinants and time course of the postthrombotic syndrome after acute deep venous thrombosis. Ann Intern Med 149: 698-707.

R5 (25) Kaklamani VG, Wisinski KB, Sadim M, Gulden C, Do A, Offit K, Baron JA, Ahsan H, Mantzoros C, Pasche B (2008) Variants of the adiponectin (ADIPOQ) and adiponectin receptor 1 (ADIPOR1) genes and colorectal cancer risk. JAMA 300: 1523-1531.

R6 (26) Kerr EA, Zikmund-Fisher BJ, Klamerus ML, Subramanian U, Hogan MM, Hofer TP (2008) The role of clinical uncertainty in treatment decisions for diabetic patients with uncontrolled blood pressure. Ann Intern Med 148: 717-727.

R7 (27) Kruijshaar ME, Watson JM, Drobniewski F, Anderson C, Brown TJ, Magee JG, Smith EG, Story A, Abubakar I (2008) Increasing antituberculosis drug resistance in the United Kingdom: analysis of National Surveillance Data. BMJ 336: 1231-1234.

R8 (28) Kuller LH, Tracy R, Belloso W, De WS, Drummond F, Lane HC, Ledergerber B, Lundgren J, Neuhaus J, Nixon D, Paton NI, Neaton JD (2008) Inflammatory and coagulation biomarkers and mortality in patients with HIV infection. PLoS Med 5: e203.

R9 (29) Laiyemo AO, Murphy G, Albert PS, Sansbury LB, Wang Z, Cross AJ, Marcus PM, Caan B, Marshall JR, Lance P, Paskett ED, Weissfeld J, Slattery ML, Burt R, Iber F, Shike M, Kikendall JW, Lanza E, Schatzkin A (2008) Postpolypectomy colonoscopy surveillance guidelines: predictive accuracy for advanced adenoma at 4 years. Ann Intern Med 148: 419-426.

R10 (30) Lederle FA, Larson JC, Margolis KL, Allison MA, Freiberg MS, Cochrane BB, Graettinger WF, Curb JD (2008) Abdominal aortic aneurysm events in the women's health initiative: cohort study. BMJ 337: a1724.

R11 (31) Limaye AP, Kirby KA, Rubenfeld GD, Leisenring WM, Bulger EM, Neff MJ, Gibran NS, Huang ML, Santo Hayes TK, Corey L, Boeckh M (2008) Cytomegalovirus reactivation in critically ill immunocompetent patients. JAMA 300: 413-422.

R12 (32) Lin GA, Dudley RA, Lucas FL, Malenka DJ, Vittinghoff E, Redberg RF (2008) Frequency of stress testing to document ischemia prior to elective percutaneous coronary intervention. JAMA 300: 1765-1773.

R13 (33) Loeb M, Hanna S, Nicolle L, Eyles J, Elliott S, Rathbone M, Drebot M, Neupane B, Fearon M, Mahony J (2008) Prognosis after West Nile virus infection. Ann Intern Med 149: 232-241.

R14 (34) Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359: 2220-2232.

R15 (35) Marcucci G, Radmacher MD, Maharry K, Mrozek K, Ruppert AS, Paschka P, Vukosavljevic T, Whitman SP, Baldus CD, Langer C, Liu CG, Carroll AJ, Powell BL, Garzon R, Croce CM, Kolitz JE, Caligiuri MA, Larson RA, Bloomfield CD (2008) MicroRNA expression in cytogenetically normal acute myeloid leukemia. N Engl J Med 358: 1919-1928.

R16 (36) McQueen MJ, Hawken S, Wang X, Ounpuu S, Sniderman A, Probstfield J, Steyn K, Sanderson JE, Hasani M, Volkova E, Kazmi K, Yusuf S (2008) Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study. Lancet 372: 224-233.

R17 (37) Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB, Sr., Cupples LA (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359: 2208-2219.

R18 (38) Meltzer ME, Lisman T, Doggen CJ, de Groot PG, Rosendaal FR (2008) Synergistic effects of hypofibrinolysis and genetic and acquired risk factors on the risk of a first venous thrombosis. PLoS Med 5: e97.

R19 (39) Mermin J, Musinguzi J, Opio A, Kirungi W, Ekwaru JP, Hladik W, Kaharuza F, Downing R, Bunnell R (2008) Risk factors for recent HIV infection in Uganda. JAMA 300: 540-549.

R20 (40) Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R, Urbauer D, Pennacchio LA, Cheng JF, Nick AM, Deavers MT, Mourad-Zeidan A, Wang H, Mueller P, Lenburg ME, Gray JW, Mok S, Birrer MJ, Lopez-Berestein G, Coleman RL, Bar-Eli M, Sood AK (2008) Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med 359: 2641-2650.

R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

(41) Montalvo G, Avanzini F, Anselmi M, Prandi R, Ibarra S, Marquez M, Armani D, Moreira JM, Caicedo C, Roncaglioni MC, Colombo F, Camisasca P, Milani V, Quimi S, Gonzabay F, Tognoni G (2008) Diagnostic evaluation of people with hypertension in low income country: cohort study of "essential" method of risk stratification. BMJ 337: a1387.
(42) Moylan CA, Brady CW, Johnson JL, Smith AD, Tuttle-Newhall JE, Muir AJ (2008) Disparities in liver transplantation before and after introduction of the MELD score. JAMA 300: 2371-2378.
(43) Nader PR, Bradley RH, Houts RM, McRitchie SL, O'Brien M (2008) Moderate-to-vigorous physical activity from ages 9 to 15 years. JAMA 300: 295-305.
(44) Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB, Sr., Kannel WB, Vasan RS (2008) A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med 148: 102-110.
(45) Peacock WF, De MT, Fonarow GC, Diercks D, Wynne J, Apple FS, Wu AH (2008) Cardiac troponin and outcome in acute heart failure. N Engl J Med 358: 2117-2126.
(46) Pearce A, Law C, Elliman D, Cole TJ, Bedford H (2008) Factors associated with uptake of measles, mumps, and rubella vaccine (MMR) and use of single antigen vaccines in a contemporary UK cohort: prospective cohort study. BMJ 336: 754-757.
(47) Peberdy MA, Ornato JP, Larkin GL, Braithwaite RS, Kashner TM, Carey SM, Meaney PA, Cen L, Nadkarni VM, Praestgaard AH, Berg RA (2008) Survival from in-hospital cardiac arrest during nights and weekends. JAMA 299: 785-792.
(48) Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Poccock S, Roberts I, Shakur H, Steyerberg E, Yutthakasemsunt S (2008) Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ 336: 425-429.
(49) Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, Overvad K, van der Schouw YT, Spencer E, Moons KG, Tjonneland A, Halkjaer J, Jensen MK, Stegger J, Clavel-Chapelon F, Boutron-Ruault MC, Chajes V, Linseisen J, Kaaks R, Trichopoulou A, Trichopoulos D, Bamia C, Sieri S, Palli D, Tumino R, Vineis P, Panico S, Peeters PH, May AM, Bueno-de-Mesquita HB, van Duijnhoven FJ, Hallmans G, Weinehall L, Manjer J, Hedblad B, Lund E, Agudo A, Arriola L, Barricarte A, Navarro C, Martinez C, Quiros JR, Key T, Bingham S, Khaw KT, Boffetta P, Jenab M, Ferrari P, Riboli E (2008) General and abdominal adiposity and risk of death in Europe. N Engl J Med 359: 2105-2120.
(50) Rawstron AC, Bennett FL, O'Connor SJ, Kwok M, Fenton JA, Plummer M, de TR, Owen RG, Richards SJ, Jack AS, Hillmen P (2008) Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. N Engl J Med 359: 575-583.
(51) Righini M, Le GG, Aujesky D, Roy PM, Sanchez O, Verschuren F, Rutschmann O, Nonent M, Cornuz J, Thys F, Le Manach CP, Revel MP, Poletti PA, Meyer G, Mottier D, Perneger T, Bounameaux H, Perrier A (2008) Diagnosis of pulmonary embolism by multidetector CT alone or combined with venous ultrasonography of the leg: a randomised non-inferiority trial. Lancet 371: 1343-1352.
(52) Ro KE, Gude T, Tyssen R, Aasland OG (2008) Counselling for burnout in Norwegian doctors: one year cohort study. BMJ 337: a2004.
(53) Sasson C, Hegg AJ, Macy M, Park A, Kellermann A, McNally B (2008) Prehospital termination of resuscitation in cases of refractory out-of-hospital cardiac arrest. JAMA 300: 1432-1438.
(54) Sattar N, McConnachie A, Shaper AG, Blauw GJ, Buckley BM, de Craen AJ, Ford I, Forouhi NG, Freeman DJ, Jukema JW, Lennon L, Macfarlane PW, Murphy MB, Packard CJ, Stott DJ, Westendorp RG, Whincup PH, Shepherd J, Wannamethee SG (2008) Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. Lancet 371: 1927-1935.
(55) Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N, Yuen ST, Chan TL, Kwong DL, Au GK, Liu CG, Calin GA, Croce CM, Harris CC (2008) MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. JAMA 299: 425-436.
(56) Schlenk RF, Dohner K, Krauter J, Frohling S, Corbacioglu A, Bullinger L, Habdank M, Spath D, Morgan M, Benner A, Schlegelberger B, Heil G, Ganser A, Dohner H (2008) Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. N Engl J Med 358: 1909-1918.
(57) Sekhri N, Feder GS, Junghans C, Eldridge S, Umaipalan A, Madhu R, Hemingway H, Timmis AD (2008) Incremental prognostic value of the exercise electrocardiogram in the initial assessment of patients with suspected angina: cohort study. BMJ 337: a2240.
(58) Smith GC, Celik E, To M, Khouri O, Nicolaides KH (2008) Cervical length at mid-pregnancy and the risk of primary cesarean delivery. N Engl J Med 358: 1346-1353.

(59) Stern DA, Morgan WJ, Halonen M, Wright AL, Martinez FD (2008) Wheezing and bronchial hyper-responsiveness in early childhood as predictors of newly diagnosed asthma in early adulthood: a longitudinal birth-cohort study. Lancet 372: 1058-1064.
(60) Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI (2008) Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 5: e165.
(61) Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med 148: 337-347.
(62) Tyson JE, Parikh NA, Langer J, Green C, Higgins RD (2008) Intensive care for extreme prematurity--moving beyond gestational age. N Engl J Med 358: 1672-1681.
(63) Tzemos N, Therrien J, Yip J, Thanassoulis G, Tremblay S, Jamorski MT, Webb GD, Siu SC (2008) Outcomes in adults with bicuspid aortic valves. JAMA 300: 1317-1325.
(64) van Veen M, Steyerberg EW, Ruige M, van Meurs AH, Roukema J, van der LJ, Moll HA (2008) Manchester triage system in paediatric emergency care: prospective observational study. BMJ 337: a1501.
(65) Vestergaard M, Pedersen MG, Ostergaard JR, Pedersen CB, Olsen J, Christensen J (2008) Death in children with febrile seizures: a population-based cohort study. Lancet 372: 457-463.
(66) Vidula H, Tian L, Liu K, Criqui MH, Ferrucci L, Pearce WH, Greenland P, Green D, Tan J, Garside DB, Guralnik J, Ridker PM, Rifai N, McDermott MM (2008) Biomarkers of inflammation and thrombosis as predictors of near-term mortality in patients with peripheral arterial disease: a cohort study. Ann Intern Med 148: 85-93.
(67) Viros A, Fridlyand J, Bauer J, Lasithiotakis K, Garbe C, Pinkel D, Bastian BC (2008) Improving melanoma classification by integrating genetic and morphologic features. PLoS Med 5: e120.
(68) Wang NC, Maggioni AP, Konstam MA, Zannad F, Krasa HB, Burnett JC, Jr., Grinfeld L, Swedberg K, Udelson JE, Cook T, Traver B, Zimmer C, Orlandi C, Gheorghiade M (2008) Clinical implications of QRS duration in patients hospitalized with worsening heart failure and reduced left ventricular ejection fraction. JAMA 299: 2656-2666.
(69) Xie J, Brayne C, Matthews FE (2008) Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up. BMJ 336: 258-262.
(70) Zethelius B, Berglund L, Sundstrom J, Ingelsson E, Basu S, Larsson A, Venge P, Arnlov J (2008) Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. N Engl J Med 358: 2107-2116.
(71) Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Gronberg H (2008) Cumulative association of five genetic variants with prostate cancer. N Engl J Med 358: 910-919

8

Principal Components Analysis to reduce the number of candidate predictors in diagnostic and prognostic prediction modeling

Nicolaas P.A. Zuithoff, MSc[1*], Mirjam I. Geerlings, PhD[1], A. Rogier T. Donders, PhD[2], Karel GM Moons, PhD[1], Yvonne Vergouwe, PhD[1],

[1]University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, the Netherlands
[2]Department of Epidemiology, Biostatistics, and HTA, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands

Submitted

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

Clinical prediction models can be used to estimate the probability of the presence of a disease (diagnosis) or the probability of a future event (prognosis)[1,2]. Examples of these models are the Framingham risk function for predicting cardiovascular disease in healthy individuals[3,4] and prediction models for the diagnosis of deep venous thrombosis in primary care[5,6]. Prediction models are commonly developed with regression analysis techniques.

The effective sample size is a key element to develop prediction models with good performance in new individuals. The effective sample size is determined by the minimum number of events or non-events for dichotomous outcomes, and the number of events with time-to-event outcomes. Simulation studies indicated that when the number of events per variable (so-called EPV) is very low, e.g. < 10, the estimates of regression coefficients and the performance of the developed prediction model become unstable[7–9]. Further, estimates have shown to be biased when backward selection is used in small samples to select the important predictors out of a much larger set of candidate predictors[2,10–13]. Regression coefficients will then be too extreme (overfitting), resulting in inaccurate predicted probabilities in future individuals[2,10–13].

Often, in studies aimed at developing a diagnostic or prognostic prediction model, the number of candidate predictors is large, certainly when compared to the number of observed events. Often employed strategies such as pre-selection based on univariable predictor-outcome associations may also lead to overfitting and biased results[14]. To prevent overfitting, it is prudent to reduce the number of candidate predictors prior to analyzing the data, if possible by utilizing the literature or clinical knowledge. Alternatively, data reduction techniques may be used that make not use of the individual predictor-outcome associations, which means that reduction is achieved by examining the associations between candidate predictors.

One of these methods is principal components analysis (PCA) in which different predictors are summarized in independent (i.e. uncorrelated) components[10,15]. Predictor reduction is then achieved by including only the most important components in a prediction model. These components still include the predictive information from several predictors, yet summarized in fewer composite variables. A thorough assessment of PCA in the development of clinical prediction models and comparison with other methods, however, is lacking.

Other methods suitable for the prevention of overfitting are Ridge and LASSO estimation[2,10,13,16–19]. These two strategies combine estimation of regression coefficients with shrinkage of the coefficients. LASSO also incorporates predictor selection by shrinking low regression coefficients to 0 [2] (Table 1).

We aimed to illustrate how to use and interpret PCA in the development of a prediction model in data with a relative low number of events per candidate predictor to reduce the number of predictors. For this aim, we used data from two empirical example studies, one from the diagnostic and one from the prognostic setting: prediction of the presence of major depressive

disorder (MDD) in primary care patients and prediction of the occurrence of cardiovascular events in patients with atherosclerotic disease. For comparison, we also developed in each example a prediction model using simple backward selection methods, Ridge regression and LASSO estimation..

Table 1. Overview of the different approaches for predictor reduction considered in this study

Approach	Predictor reduction	Shrinkage incorporated*
PCA	±	-
Full model	-	-
Backward selection	+	-
Ridge regression	-	+
LASSO	+	+

+ Applicable
- Not applicable
± Predictor reduction is not inherent to PCA, but rather applied to reduce the number of coefficients to be estimated.
* for approaches that do not incorporate shrinkage, internal validation was used to calculate a shrinkage factor (see text for details).

Methods

Example studies

To obtain generalisable results, we selected one example of a diagnostic study and one example of a prognostic study from different clinical domains.

Diagnosis of major depressive disorder (MDD)

For this example we used data from the Dutch part of the PredictD study (PREDICT-NL). PredictD is a large international prospective cohort study from which a multifactor algorithm was developed to predict the onset of MDD in primary care patients[20]. The design and primary results from the PredictD study can be found elsewhere[20,21]. Information of 1046 patients from the PREDICT-NL study was used for the development of a clinical prediction model for estimating the probability of MDD presence[22]. The presence or absence of MDD was assessed in all patients according to the DSM-IV criteria.

Prior to the analysis, the data was split non-randomly by geographical area, to mimic the development of prediction models and externally validate their predictive accuracy in patients from another place and institute: geographical validation[23–25].

We here considered 22 candidate predictors for the model for MDD (Appendix 1), collected with questionnaires and from patients medical records. The candidate predictors included 11 'life events' that were derived from a questionnaire developed by Brugha et. al.[26], in addition to 11 other predictors (Appendix 1). In the development data, we analysed 77 events with 22 candidate predictors, or 3.5 events per variable. Hence, overfitting of the models was very likely.

Prognosis of patients with atherosclerotic disease

For this example we used data from the SMART study (Second Manifestations of ARTerial disease). The SMART study is a prospective cohort study among patients with either symptomatic atherosclerotic disease (e.g. coronary heart disease, cerebrovascular disease) or risk factors for atherosclerosis (e.g. hypertension, diabetes mellitus). The details of this study were previously reported[27]. Briefly, at baseline, all patients underwent a standardized vascular screening including laboratory measurements and ultrasonography and filled in a questionnaire on health and lifestyle. In the follow up, all new cardiovascular events were registered. For this study, we analyzed a composite endpoint of fatal or non-fatal coronary ischemic events, fatal or non-fatal ischemic stroke and vascular death.

For this paper, we specifically selected patients that were included between jan 1st 2000 and dec 31st 2000 for model development. Data of patients included later were used for external validation: temporal validation[23–25].

We considered 14 candidate predictors for the model for cardiovascular events. The candidate predictors included 10 continuous measurement ('physical measurements') such as body mass index (BMI) and lipid values, in addition to four other known risk factors for cardiovascular events (Appendix 3). As BMI had a U-shaped association with the risk of cardiovascular events, this variable was transformed (i.e. setting the mean of the distribution to 0 and squared) prior to analysis. In the development data, we analysed 75 cardiovascular events with 14 predictors, or 5.4 events per variable. Hence, overfitting of models was again likely.

Approaches for predictor reduction and model development

Principal Component Analysis (PCA)

A thorough description of PCA can be found elsewhere[28]. In brief, with PCA, new variables (components) are constructed as weighted summations of the original variables, based on correlations[28,29]. Any of the original variables included in the PCA will contribute ("load") to all the components, although the weight ("loading") for the same variable will differ for different components. The weights of the original variables on each component are calculated in such manner that the first component explains the most variance of the original variables whereas any subsequent component explains any variance not yet captured in preceding components. As a result, components are ordered by the amount of information retained from the original variables. PCA will initially extract as many components as the number of variables originally included. Reduction of the number of regression coefficients to be estimated in model development is achieved by selecting components that explain the most variance[15]. For this study, we selected components that retain more variance than one original variable[30]. The selected components were included in the models rather than the original predictors. As we performed PCA based on correlations, predictors selected for PCA were effectively standardized (i.e. setting the mean to 0 and the standard deviation to 1). We therefore standardized these predictors in all analysis.

In the MDD example, we specifically selected the 11 predictors of 'life events' as there was considerable overlap in content between these items in the questionnaire.
In the cardiovascular example, we explicitly selected all continuous measurements (except age), as we assumed these physical measurements could effectively summarized in fewer components.

Backward selection

In this approach we started in each example with the full model (i.e including all candidate predictors) and reduced the model by eliminating statistically non-significant predictors (using a p-value >0.05 based on the likelihood ratio test). This approach was compared as it is one of the most often applied types of predictor selection. As it is known that such backwards predictor selection using p-values at he nominal level of 0.05 may result in overestimation of selected regression coefficients and optimistic performance, we also applied Akaike's Information Criterion (AIC) as selection criterion for predictor retainment in the model[10,15]. AIC is equal to a p-value of 0.157, if a predictor is modelled with one regression coefficient.

Ridge estimation

Ridge estimation is a form of penalized maximum likelihood estimation (PMLE)[17]. The 'standard' maximum likelihood function for the estimation of the regression coefficients is directly (during model fitting) adjusted by a so-called penalty factor. Shrinkage of regression coefficients is in this way directly incorporated in the model fitting process Ridge estimation fits a full model without predictor selection and applies a different amount of shrinkage for each predictor. A more thorough description of Ridge estimation can be found elsewhere[17].

LASSO estimation

LASSO (least absolute shrinkage and selection operator) estimation is, like Ridge estimation, a form of PMLE with an inherent adjustment in the model fitting prcoes of the maximum likelihood function by a penalty factor (see Goeman[16] for details). LASSO estimation is applied to a full model. Strong predictors will have comparatively little shrinkage whereas weak predictors will shrink to exactly 0. LASSO therefore has the additional advantage of selection of predictors. A thorough description of LASSO can be found elsewhere[16,19,31].

Performance measures

The performance of the models was assessed in terms of discrimination and calibration. Discrimination, the ability of a models to distinguish between patients with and without a specific event, was quantified with the concordance-statistic (c-statistic)[32]. Calibration, the agreement between predicted probabilities and observed frequencies, was quantified with a calibration slope[10,15]. The calibration slope was calculated by multiplying the predictors of individual patients with the regression coefficients from the developed models: the thus derived linear predictor

was subsequently analyzed as a single predictor. Models with perfect calibration should have a calibration slope of 1.

Internal and external validation

Internal validation of the predictive performance of the developed models was carried out with bootstrapping techniques[2,10,15,33]. In every randomly drawn bootstrap sample, the entire modelling process, including either PCA, backward selection or the PMLE estimations, was repeated. The resulting model was then applied to the original sample to calculate the predicted risks that were compared to the observed outcomes. This process was repeated 100 times to obtain a stable estimate of the calibration slope and the optimism adjusted c-statistic.

External validity of the models could be assessed on patients from different general practices (MDD) or treated more recently (cardiovascular disease). For the methods that do not include shrinkage (full models, PCA, backward selection), we used the calibration slope from the internal validation as a shrinkage factor[2,10,15,33]. The performance of all models was quantified with the calibnration slope and the c-statistic.

All analysis were performed with R (version 2.11.1; R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org. We used Harrell's rms package and Goeman's penalized package[16].

Results

Diagnosis of major depressive disorder (MDD)

Model development

PCA reduced the 11 original life event predictors to four components (Appendix 2). We then included these four components as predictors in the model together with the 11 other predictors consequently modelling the presence of MDD with 15 instead of 22 predictors. The model developed with backward selection based on AIC retained four of the 11 original life event predictors (Table 2). The model developed with LASSO estimation selected only two of the four life events that were retained after the backward selection approach. Both models that included all predictors, the full model and the model developed with Ridge estimation, had the highest c-statistic of 0.87 in the development set (Table 3). All other models had similar discriminative ability of 0.84 to 0.85.

Table 2. Logistic Regression coefficients in the developed models for the diagnosis of major depressive disorders

	PCA	Full model	Backward selection AIC	Backward selection p<.05	Ridge	LASSO
Female gender	0.31	0.30	-	-	0.27	0.09
Age, years	-0.01	-0.01	-	-	-0.02	-0.02
Educational level (None/primary only)	0.37	0.38	0.36	-	0.37	0.23
Being single	-0.07	-0.23	-		-0.07	-
Number of presented complaints						
2	0.22	0.14	-	-	0.14	-
≥3	0.63	0.63	-	-	0.44	-
Non-somatic diagnosis/complaint at inclusion consult	-0.01	0.03	-	-	0.07	-
Consultation rate (number of consults in the past 12 months)	-0.01	-0.005	-	-	0.01	0.02
Received depression codes in past 12 months	1.79	1.79	2.10	2.26	1.14	1.72
Prescription of antidepressants in the past 12 months	0.53	0.38	-	-	0.70	0.40
Any depressed feelings, lifetime	0.79	0.83	0.87	0.94	0.85	1.00
Any loss of interest, lifetime	0.61	0.55	0.58	0.60	0.65	0.54
Predictors included in PCA						
Serious illness, injury or assault, self	x	0.37	-	-	0.19	-
Serious illness, injury or assault, close relative	x	-0.01	-	-	0.06	-
Death of parent, child or spouse	x	0.77	0.55	0.65	0.55	0.34
Death of friend or distant relative	x	-0.31	-	-	-0.16	-
Separation due to marital difficulties	x	-1.57	-1.53	-1.56	-0.31	-
End of steady relationship	x	1.12	1.26	1.31	0.40	-
Serious problem with close friend, neighbour or relative	x	0.77	0.86	0.90	0.73	0.85
Unemployed or seeking job unsuccesfully	x	0.17	-	-	0.26	-
Sacked from job	x	-0.05	-	-	-0.03	-
Major financial crisis	x	0.19	-	-	0.21	-
Something valuable stolen or lost	x	0.30	-	-	0.16	-
Components from PCA						
Component 1	-.12	x	x	x	x	x
Component 2	0.04	x	x	x	x	x
Component 3	0.09	x	x	x	x	x
Component 4	0.02	x	x	x	x	x

Note: Coefficients were shrunken, see methods section for details
- Not selected
x Only implicitely considered in the model

Internal Validation

The calibration slope estimated by internal validation showed the best results, i.e. a value closest to 1, for Ridge estimation (Table 3). Internal validaton yielded the optimism adjusted c-index which was highest for the full model and for both models developed with backward selection (Table 3).

External validation

In the external validation set, the calibration slope for the model developed with PCA had a value of 0.86 (Table 3). By comparison, the slope for backward selection with AIC was 0.74, whereas the slope for backward selection with p<0.05 was only 0.68. Ridge estimation showed the best external calibration, with a slope of 0.93. The slope for LASSO estimation was 0.85.

The c-statistic in the validation set of the model developed with PCA was 0.77. Somewhat lower values of 0.73 to 0.75 were observed for the full model and for both models developed with backward selection. The models developed with Ridge and LASSO estimation had very similar values of 0.77 and 0.78, respectively.

Table 3. Performance of the different fitted models for major depressive disorder

	Calibration Slope		Concordance-statistic		
	Internal validation	External validation	Apparent	Internal validation	External validation
Model with PCA	0.78	0.86[1]	0.85	0.77	0.77
Full model	0.69	0.74[1]	0.87	0.81	0.75
Model after backward selection with AIC	0.71	0.74[1]	0.85	0.80	0.74
Model after backward selection with p<.05	0.76	0.68[1]	0.86	0.81	0.73
Model with Ridge estimation	0.81	0.93	0.87	0.74	0.78
Model with LASSO estimation	0.80	0.85	0.84	0.73	0.77

[1] The calibration slope from internal validation was used to shrink regression coefficients before calculating the calibration slope in the external validation data

Prognosis of patients with atherosclerotic disease

Model development

The PCA of physical measurements was used to reduce the 10 original predictors to five components (Appendix 4). We included the five components in a prediction model for cardiovascular events, consequently analyzing cardiovascular events with nine instead of 14 predictors. The model developed with backward selection with AIC retained four of the original predictors: HDL cholesterol, hemoglobine, systolic bloodpressure, HbA1c and intima-media thickness (IMT), as well as age (Appendix 4). When we applied the more stringent p-value of 0.05 for selection, HDL cholesterol was also excluded as a predictor. The model developed with LASSO estimation retained age, fasting glucose, HDL cholesterol, HbA1c and IMT, shrinking all other

coefficients to zero. In Ridge estimation, all predictors were retained in the model. Gender, age, HDL cholesterol and IMT were the strongest predictors. All models had very similar c-statistics of 0.77, only the c-statistic of the model developed with LASSO had somewhat lower value of 0.75.

Table 4. Cox regression coefficients for the differently fitted cardiovascular models

	PCA	Full model	Backward selection AIC	Backward selection p<.05	Ridge	LASSO
Male gender	0.87	0.95	0.83	0.99	0.25	-
Age, per 10 years	0.33	0.28	0.30	0.30	0.45	0.51
Diabetes present	0.08	0.02	-	-	0.05	-
Smoking						
Never	0	0	-	-	-0.06	-
Former	0.11	0.10	-	-	0.03	-
Present	0.10	0.10	-	-	0.03	-
Predictors included in PCA						
Fasting glucose*	x	0.07	-	-	0.09	0.03
BMI squared*	x	-0.03	-	-	-0.05	-
Hemoglobine*	x	-0.20	-0.19	-0.18	-0.11	-
Triglyceride*	x	0.007	-	-	-0.02	-
HDL cholesterol*	x	-0.20	-0.23	-	-0.24	-0.20
LDL cholesterol*	x	0.14	-	-	0.04	-
Systolic bloodpressure*	x	0.31	0.19	0.20	0.14	-
Diastolic bloodpressure*	x	-0.21	-	-	-0.06	-
HbA1c*	x	0.16	0.22	0.23	0.12	0.08
Intima-media thickness*	x	0.19	0.21	0.22	0.24	0.26
Components from PCA						
Component 1	0.18	x	x	x	x	x
Component 2	-0.10	x	x	x	x	x
Component 3	0.13	x	x	x	x	x
Component 4	-0.32	x	x	x	x	x
Component 5	-0.05	x	x	x	x	x

Note: Coefficients were shrunken, see methods section for details
- Not selected
x Not considered in the model
* per SD

Internal Validation

The calibration slope estimated from internal validation for the PCA model was 0.88 (Table 5). However, the calibration slopes showed the best results for LASSO estimation with a value of 1.00. The 'second best' value from internal validation was observed for Ridge estimation with values of 0.95. Internal validaton yielded the optimism adjusted c-index which showed some variation between 0.72 and 0.76. The optimism adjusted c-index was highest for the model with PCA.

External validation

In the external validation set, the calibration slope for the model developed with PCA had a value of 0.80, whereas the models developed with backward selection showed slopes of 0.94 ($p<0.05$) and 0.95 (AIC). LASSO and Ridge estimation showed values of 0.94 and 0.88 respectively. The best calibration was observed for models with the least number of predictors.

The c-statistic in the validation set was very similar for all models, with values around 0.70 (Table 5). In an additional analysis (results not shown) we observed that age was the strongest predictor (achieving an apparent c-statistic of 0.69), whereas models without age showed a dramatic decrease in performance in external validation.

Table 5. Performance of the different fitted models for cardiovascular events

	Calibration Slope		Concordance-statistic		
	Internal validation	External validation	Apparent	Internal validation	External validation
Model with PCA	0.88	0.80[1]	0.77	0.75	0.71
Full model	0.82	0.86[1]	0.78	0.74	0.69
Model after backward selection with AIC	0.81	0.95[1]	0.77	0.74	0.70
Model after backward selection with p<.05	0.83	0.94[1]	0.77	0.73	0.69
Model with Ridge estimation	0.95	0.92	0.77	0.74	0.70
Model with LASSO estimation	1.00	0.98	0.75	0.72	0.70

[1] The calibration slope from internal validation was used to shrink regression coefficients before calculating the calibration slope in the external validation data

Discussion

We compared different variable reduction techniques when developing a prediction model in situatons where the number of candidate predictors is high, with a specific focus on PCA as compared to backward selection, Ridge and LASSO estimation. The four methods deal differently with two important issues in prediction model development: i) predictor selection; ii) shrinkage of estimated regression coefficients. In the MDD example, we observed good external calibration and discrimination for models developed with PCA but also with LASSO and Ridge estimation. A full model and the models developed with backward selection showed poorer calibration in new patients. In the cardiovascular events example, the results were driven by one dominant predictor. As a consequence, the discriminative ability in the new patients was very similar for all models. The model with PCA showed the worst calibration, whereas the best calibration was shown with LASSO estimation.

In the MDD example, we developed models with low numbers of events per variable. The relative poor performance of backward selection strategies and the full model under this condition, even after applying shrinkage, has been well established[2,10,15,33]. Consistent with statistical

recommendations[2,10,15,16,33], models developed with LASSO and Ridge performed substantially better. The model developed with PCA also showed good performance. With PCA, we achieved a reduction of the number of regression coefficients without much loss of predictive information and without analyzing the association with the outcome. Contrary to guidelines, models for cardiovascular events that were developed with backward selection showed good discrimination at external validation. This was due to the fact that age was the most important predictor. The information of the physical measurements was summarized with PCA, the additional predictive value of the physical measurements, however, was limited.

The question may arise when PCA is a suitable method for the reduction of the number of predictors in a prediction study. First, we note that PCA, together with backward selection and LASSO and Ridge estimation, is essentially a data-driven method. Preselection of predictors based on substantial external knowledge (i.e. expert opinion) is preferred over data-driven methods. If such preselection is feasible, there is no need to employ PCA for the reduction of the number of predictors. Second, components derived from PCA are always a composite of the included predictors. As such, PCA is not suitable to determine the predictive value of one specific predictor or to select the best predictors from a group of candidate predictors. Consequently, PCA may be valuable in prediction studies where the aim is to construct a stable prediction model that produces accurate probabilities of an event of interest, rather than the selection of important predictors or the evaluation of a specific predictor.

A possible advantage of PCA over Ridge and LASSO estimation may be the relatively straightforward approach. PCA is available in almost all standard software for data analysis (SPSS, SAS, Stata, R, S-plus). The components derived from PCA are linear summations of the original predictors. In the model fitting process, each component is considered as a predictor. The following prediction model can be written as a model that contains the same predictors as the full model. The weights are now based on the regression coefficients and the weights from the PCA. An example is given in Appendix II.

We applied PCA on a selection of candidate predictors, based on overlap between items (i.e. life event items in MDD example) or similar measurement level (i.e. continuous predictors in the cardiovascular events example). In the subsequent models, we included only the components that included more information than one original variable. Other choices may lead to different results. For example, the number of components could be determined based on the choice of EPV, which implies a maximum number of regression coefficients given the number of events. A requirement could also be that the components included in the prediction model should contain a fixed minimum of explained variance of the original variables. Simulation studies may be employed to study the consequences of different criteria.

This study has some limitations. First, we presented the results of two case studies. As such, some findings may be specific, although realistic, for the examples presented here. Second, the choice of predictors that are included in the PCA may influence the results. We choose groups of predictors

that, from a clinical perspective, were related and could therefore reasonably be described by fewer components. Third, in the cardiovascular example, we assumed linear associations between the outcome and all continuous predictors, except BMI. BMI was analyzed as a squared predictor. Reduction of the number of predictors in the presence of non-linear associations between predictors and outcomes may require more advanced forms of PCA[15].

In conclusion, when developing a diagnostic or prognostic prediction model, we have illustrated the use of PCA for predictor reduction in situations where the number of candidate predictors is high relative to the number of events. In the absence of a dominant predictor, PCA seems preferable over ordinary backward selection methods, and is as good as Ridge regression and LASSO techniques to deal with large numbers of candidate predictors.

References

(1) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338: b375.

(2) Steyerberg, E. W. (2009) Clinical prediction models; a practical approach to development, validation, and updating. New York: Springer.

(3) Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837-1847.

(4) Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS (2009) Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation 119: 3078-3084.

(5) Oudega R, Moons KG, Hoes AW (2005) Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 94: 200-205.

(6) Wells PS, Owen C, Doucette S, Fergusson D, Tran H (2006) Does this patient have deep vein thrombosis? JAMA 295: 199-207.

(7) Concato J, Peduzzi P, Holford TR, Feinstein AR (1995) Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 48: 1495-1501.

(8) Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48: 1503-1510.

(9) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379.

(10) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(11) Steyerberg EW, Eijkemans MJ, Habbema JD (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 52: 935-942.

(12) Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 56: 441-447.

(13) Moons KG, Donders AR, Steyerberg EW, Harrell FE (2004) Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 57: 1262-1270.

(14) Sun GW, Shook TL, Kay GL (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 49: 907-916.

(15) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(16) Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. Biom J 52: 70-84.

(17) Le Cessie S, Van Houwelinge JC (1992) Ridge Estimators in Logistic Regression. Applied Statistics 41: 191-201.

(18) Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.

(19) Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society 58: 267-288.

(20) King M, Weich S, Torres F, Svab I, Maaroos H, Neeleman J, Xavier M, Morris R, Walker C, Bellon JA, Moreno B, Rotar D, Rifel J, Aluoja A, Kalda R, Geerlings MI, Carraca I, Caldas dA, Vincente B, Salidivio S, Rioseco P, Nazareth I (2006) Prediction of depression in European general practice attendees: the PREDICT study. BMC Public Health 6: 6.

(21) King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, Bellon-Saameno JA, Moreno B, Svab I, Rotar D, Rifel J, Maaroos HI, Aluoja A, Kalda R, Neeleman J, Geerlings MI, Xavier M, Carraca I, Goncalves-Pereira M, Vicente B, Saldivia S, Melipillan R, Torres-Gonzalez F, Nazareth I (2008) Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study. Arch Gen Psychiatry 65: 1368-1376.

(22) Zuithoff NP, Vergouwe Y, King M, Nazareth I, Hak E, Moons KG, Geerlings MI (2009) A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Fam Pract 26: 241-250.

(23) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.

(24) Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med 130: 515-524.
(25) Toll DB, Janssen KJ, Vergouwe Y, Moons KG (2008) Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 61: 1085-1094.
(26) Brugha TS, Cragg D (1990) The List of Threatening Experiences: the reliability and validity of a brief life events questionnaire. Acta Psychiatr Scand 82: 77-81.
(27) Simons PC, Algra A, van de Laak MF, Grobbee DE, van der GY (1999) Second manifestations of ARTerial disease (SMART) study: rationale and design. Eur J Epidemiol 15: 773-781.
(28) Jolliffe, I. T. (1986) Principal Components Analysis. New York: Springer-Verlag.
(29) Dunteman, G. H. (2011) Prinicpal Components Analysis. Newbury Park, Ca: Sage Publications Ltd.
(30) Kaiser HF (1960) The application of electronic computers to factor analysis. Educational and Psychological measurement 20: 141-151.
(31) Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.
(32) Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247: 2543-2546.
(33) Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54: 774-781.

Appendix 1

Table Characteristics of patients in the PREDICT-NL study

	Development (N=558)	Validation (N= 486)
Candidate predictors		
Female gender	63 (352)	66 (321)
Age, years[1]	46 (13)	44(13)
Educational level (None/primary only)	25 (139)	16 (76)
Being single	20 (112)	26 (127)
Number of presented complaints		
1	79 (439)	79 (386)
2	16 (92)	15 (74)
≥3	5 (27)	6 (28)
Non-somatic diagnosis/complaint at inclusion consult	9 (50)	8 (37)
Consultation rate (number of consults in the past 12 months)[2]	9 (5-14)	8 (5-13)
Received depression codes in past 12 months	4 (24)	7 (34)
Prescription of antidepressants in the past 12 months	8 (44)	9 (46)
Any depressed feelings, lifetime	46 (259)	52 (253)
Any loss of interest, lifetime	38 (210)	43 (207)
Predictors included in PCA		
Life events in the past 6 months		
Serious illness, injury or assault, self	5 (30)	7 (34)
Serious illness, injury or assault, close relative	20 (113)	19 (94)
Death of parent, child or spouse	15 (81)	11 (55)
Death of friend or distant relative	29 (160)	27 (132)
Separation due to marital difficulties	3 (17)	5 (23)
End of steady relationship	4 (22)	7 (32)
Serious problem with close friend, neighbour or relative	12 (67)	12 (58)
Unemployed or seeking job unsuccesfully	5 (27)	8 (39)
Sacked from job	3 (19)	5 (24)
Major financial crisis	6 (33)	10 (47)
Something valuable stolen or lost	5 (29)	6 (29)
Outcome		
Major depressive disorder	14 (77)	17 (82)

Characteristics are reported als % (n), except where noted otherwise
[1] Mean (SD)
[2] Median (interquartile range)

Appendix 2

Tabel Component loadings from PCA of Life Events

	Component 1	Component 2	Component 3	Component 4
Serious illness, injury or assault, self	-0.164	-0.101	0.489	-0.263
Serious illness, injury or assault, close relative	-0.231	0.343	0.290	0.091
Death of parent, child or spouse	-0.180	0.537	-0.259	-0.195
Death of friend or distant relative	-0.245	0.550	-0.035	-0.058
Separation due to marital difficulties	-0.370	-0.225	-0.258	-0.493
End of steady relationship	-0.361	-0.372	-0.035	-0.391
Serious problem with close friend, neighbour or relative	-0.353	-0.019	0.331	0.275
Unemployed or seeking job unsuccesfully	-0.409	-0.148	-0.194	0.440
Sacked from job	-0.386	-0.069	-0.437	0.319
Major financial crisis	-0.334	-0.018	0.449	0.073
Something valuable stolen or lost	-0.092	0.254	-0.029	-0.329
Explained variance (%)	19	14	11	10

Appendix 3

Table Characteristics of patients in the SMART study

	Development (N=528)	Validation (N= 4040)
Candidate predictors		
Male gender, %(n)	67 (352)	67 (2720)
Age, years	56 (13)	55 (13)
Diabetes present, %(n)	18 (97)	21 (863)
Smoking		
Former, %(n)	45 (237)	43 (1752)
Present, %(n)	31 (166)	30 (1221)
Predictors included in PCA		
Fasting glucose (mmol/l)	6.5 (2.6)	6.4 (2.2)
BMI (kg/m^2)	27 (4.3)	27 (4.5)
Hemoglobine (mmol/l)	9.0 (0.8)	8.9 (0.8)
Triglyceride (mmol/l)	2.1 (1.6)	1.9 (2.0)
HDL cholesterol (mmol/l)	1.2 (0.4)	1.3 (0.4)
LDL cholesterol (mmol/l)	3.5 (1.1)	3.0 (1.2)
Systolic bloodpressure (mmHg)	139 (23)	143 (21)
Diastolic bloodpressure (mmHg)	81 (12)	85 (12)
HbA1c (%)	6.3 (1.2)	6.1 (1.0)
Intima-media thickness (mm)	0.89 (0.30)	0.88 (0.28)
Outcome		
Cardiovascular events, %(n)	14 (75)	6 (228)
Time, years, median (IQR)[1]	7.5 (7.2-7.9)	1216 (1.6-5.0)

Characteristics are expressed as means (SD), except where noted otherwise
[1] IQR – Inter Quartile Range

Appendix 4

Table PCA of physical measurements in the SMART study

	Component 1	Component 2	Component 3	Component 4	Component 5
Fasting glucose	0.63	-0.05	0.15	-0.05	0.27
BMI squared	0.17	-0.04	0.18	-0.34	-0.33
Hemoglobine	-0.08	-0.13	-0.44	-0.11	0.57
Triglyceride	0.19	-0.39	-0.47	-0.00	-0.21
HDL cholesterol	-0.18	0.21	0.51	-0.37	0.09
LDL cholesterol	-0.24	0.03	0.18	0.23	0.60
Systolic bloodpressure	-0.06	-0.62	0.35	0.08	-0.03
Diastolic bloodpressure	-0.15	-0.62	0.17	-0.27	0.14
HbA1c	0.64	0.03	0.15	-0.02	0.24
Intima-media thickness	0.06	-0.10	0.25	0.77	-0.12
Explained variance (%)	19	17	15	12	11

Appendix 5.

In this example we will show how the components from the cardiovascular example can be included in the prediction model. Initially, we will only describe the incorporation of the first component from PCA, ignoring the other 4 components. The results for all components are given in the table below. The physical measurements were standardized (i.e. setting the mean of the distribution at zero and the SD at 1) in the PCA. Prior to calculating the score, the measurements should be transformed by subtracting the mean of the distribution and dividing this by the standard deviation. Standardized values are denoted here by the Z subscript. The values are given in Table 3 in Appendix I. The first component score (CS1) of the physical measurements is given by:

(1) CS1 = 0.63 * Fasting glucose$_z$ + 0.17 * BMI squared$_z$ + 0.06 * Hemoglobine$_z$ + 0.19 * Triglyceride$_z$ + -.18 * HDL cholesterol$_z$ + -.24 * LDL cholesterol$_z$ + -.06 * Systolic bloodpressure$_z$ + -.15 * Diastolic bloodpressure$_z$ + 0.64 * HbA1c$_z$ + 0.06 * Intima-media thickness$_z$

The prediction model is given in Table 4 (incorporating only the first component). The linear predictor (LP) of the model is given by:

(2) LP = 0.87 * Male gender + 0.33 * Age, per 10 years + 0.08 * Diabetes present + 0.11 * Former smoking+ 0.10 * Now smoking + 0.18 * CS1

Formula (1) from the PCA can now simply be incorporated in formula (2):

(3) LP = 0.87 * Male gender + 0.33 * Age, per 10 years + 0.08 * Diabetes present + 0.11 * Former smoking+ 0.10 * Now smoking + 0.18* (0.63 * Fasting glucose$_z$ + 0.17 * BMI squared$_z$ - 0.08 * Hemoglobine$_z$ + 0.19 * Triglyceride$_z$ + -.18 * HDL cholesterol$_z$ + -.24 * LDL cholesterol$_z$ + -.06 * Systolic bloodpressure$_z$ + -.15 * Diastolic bloodpressure$_z$ + 0.64 * HbA1c$_z$ + 0.06 * Intima-media thickness$_z$)

Recalculation of the PCA part of the formula gives:

(4) LP = 0.87 * Male gender + 0.33 * Age, per 10 years + 0.08 * Diabetes present + 0.11 * Former smoking + 0.10 * Now smoking + 0.11 * Fasting glucose$_z$ - 0.03 * BMI squared$_z$ - 0.13 * Hemoglobine$_z$ + 0.02 * Triglyceride$_z$ – 0.11 * HDL cholesterol$_z$ -.02 * LDL cholesterol$_z$ - 0.12 * Systolic bloodpressure$_z$ - 0.04 * Diastolic bloodpressure$_z$ + 0.12 * HbA1c$_z$ + 0.31 * Intima-media thickness$_z$

The other components can be included the same way as described for the first component. The inclusion of the other components will lead to 5 terms for each physical measurements in formula (4) that can simply be added to one weight for each measurement.

For a prediction model intended for use in clinical practice, the weights of the physical measurements in the PCA could subsequently be recalculated to the original scales of the measurements by multiplying them with their respective standard deviations. Furthermore, the probabilities for cardiovascular events can be calculated by relating the linear predictor of the model to the baseline risks from the Cox regression model.

Table. Overview of the calculation of the final coefficients for each predictor from PCA and regression analysis.

	Component 1	Component 2	Component 3	Component 4	Component 5	Final Coefficient
	Loading * regression coefficient	Loading * regression coefficient	Loading * regression coefficient	Loading * regression coefficient	Loading * regression coefficient	
Regression coefficient for components	0.19	-0.10	0.13	0.32	-0.06	
Fasting glucose	0.12	0.00	0.02	-0.02	-0.02	0.11
BMI squared	0.03	0.00	0.02	-0.11	0.02	-0.03
Hemoglobine	-0.01	0.01	-0.06	0.04	-0.03	-0.13
Triglyceride	0.03	0.04	-0.06	-0.00	0.01	0.02
HDL cholesterol	-0.03	-0.02	0.07	-0.12	-0.01	-0.11
LDL cholesterol	-0.04	-0.00	0.02	0.07	-0.04	0.02
Systolic bloodpressure	-0.01	0.06	0.05	0.03	0.00	0.12
Diastolic bloodpressure	-0.03	0.06	0.02	0.09	-0.01	-0.04
HbA1c	0.12	0.00	0.02	-0.01	-0.01	0.12
Intima-media thickness	0.01	0.01	0.03	0.25	0.01	0.31

Note: all coeffcients are valid for standardized values of the predictors. Original loadings are given in Appendix 4.

9

Dichotomising continuous outcomes in prediction research, a bad idea?

Nicolaas P.A. Zuithoff, MSc[1*], Mirjam I. Geerlings, PhD[1], A. Mireille Baart[1,2], Wim L.A.M. de Kort, PhD[2], Karel G.M. Moons, PhD[1], Yvonne Vergouwe, PhD[1]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Sanquin Blood Supply, Sanquin Research, Dept. Donor Studies, Nijmegen, the Netherlands

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Introduction

Clinical prediction models usually estimate the probability of the presence of having a a certain disease (diagnosis) or developing a future event (prognosis)[1]. Examples of these models are the Framingham risk function for predicting the occurrence of cardiovascular disease in healthy individuals[2,3] and prediction models for the diagnosis of deep venous thrombosis in primary care patients[4,5].

In epidemiological research, continuous measurements are frequently dichotomised. For example, blood pressure values are dichotomised to define patients as hypertensive or non-hypertensive. Although dichotomisation of continuous measurements may seem helpful in clinical practice to make decisions, it commonly leads to loss of information, and therefore loss of statistical power and increased chance of type I error[6–8]. Further, it may introduce spurious associations and the results depend on the specific cutoff levels used for dichotomisation[7,9]. Dichotomisation of continuous predictor variables in the analysis process has been labelled as 'a bad idea'[10].

The disadvantages of dichotomising predictor variables may also apply to dichotomising continuous outcome variables. For example, risk prediction of hypertension may well be replaced by predicting the absolute blood pressure value. These continuous outcomes are usually modelled with linear regression, whereas models with dichotomous outcomes are usually developed with logistic regression analysis[11]. Even though disadvantages of dichotomising continuous outcomes has previuously been studied[6,9,12–15], consequences for clinical prediction models have never been evaluated.

The development of a prediction model with a continuous outcome may especially be advantageous in small study samples, due to the effective sample size. The effective sample size for continuous outcomes is equal to the sample size, the total number of study participants. For a dichotomous outcome, however, the effective sample size is equal to the smallest of the number of outcome events or non-events[11]. Prediction models developed in data with a small effective sample size and a relative large number of predictors, for dichotomous outcomes often quantified as events-per-variable (or EVP), may be too optmistic (overfitted)[11,16]. Prediction models developed for a continuous outcomes may thus give better performance (less overfitting) in new patients.

In this study, we aimed to compare the performance of linear prediction models that predict the underlying continuous outcomes versus logistic prediction models for predicting the dichotomised version of such continuous outcomes. This is done using the empirical data of two example studies. In the first example, we compare the prediction of the presence of high score on the PHQ-9 depression scale (indicative of clinically relevant depression) as compared with the prediction of the continuous PHQ-9 depression score. In the second example, the prediction of the presence of low hemoglobin level (Hb deferral) in blood donors is compared with the prediction of the absolute hemoglobin level.

Methods

Empirical data

The first example is the prediction of major depressive disorder using the data of the previously published PREDICT-NL study. In PREDICT-NL, information of 1046 patients visiting their primary care physician was used for the development of a prediction model for estimating the risk of having major depressive disorder[17]. The design and primary results from this study can be found elsewhere[18,19]. For the purpose of the present study, we used the original continuous depression scores from the Patient Health Questionnaire-9 (PHQ-9) as outcome[20,21]. PHQ-9 scores of 6 or higher are considered as clinically relevant depression[22].

Prior to the analysis, the data was split non-randomly (by general practices) in a development set (n=547) and a validation set (n=499). A non-random split was specifically done, as it reduces the similarities between the development and validation data sets[23,24], thus introducing a more rigorous assessment of the external validity. Prediction models were developed for the continuous and for the dichotomised PHQ-9 score. We considered 12 candidate predictors for both outcomes (Appendix 1). In the logistic model, we analysed 146 cases of depression with these 12 predictors, hence some overfitting was likely.

The second example is the prediction of low hemoglobin (Hb) levels, or Hb deferral. Female blood donors who visited a blood collection center in the Netherlands between 2007 and 2009 were included in the study. The design and primary results from this study can be found elsewhere[25]. Hb levels were dichotomised at the accepted cutoff value of 7.8 mmol/L for women. Prediction models were developed for the continuous and for the dichotomised Hb values. We considered 10 candidate predictors for both outcomes (Appendix 2). Data of 35,687 donors was used for model development; data of 29,169 donors from another region were used for model validation. In the development data, 3033 cases of low Hb were analysed. Hence, overfitting of the logistic model was not expected.

The development of a prediction model with a continuous outcome may especially be advantageous in small study samples. The effective sample size for continuous outcomes is equal to the sample size, the total number of study participants. For a dichotomous outcome, however, the effective sample size is equal to the smallest of the number of outcome events or non-events[11]. As the Hb deferral example is based on a large development sample, we also created a small subsample for model development. We randomly selected 765 participants (65 events) from the development data to develop models with 5 events per variable for the dichotomous outcome, introducing a potentially high degree of overfitting for the logistic model.

Model development

We developed prediction models with regression analysis. Continuous outcomes were analyzed with linear regression. Dichotomised outcomes were analyzed with logistic regression. For the

continuous and dichotomised outcomes, we first developed the corresponding full models, i.e. including all candidate predictors. We subsequently reduced the models with backward selection based on likelihood ratio tests. In the complete Hb development sample, we applied the standard p-value of 0.05, as the effective sample size for both outcomes was very large. In the small development samples (depression and Hb subsample), we used Akaike's Information Criterion (AIC) as stopping rule[11,16].

Performance measures

The performance of the linear and logistic model, in both examples, was assessed in terms of calibration, the agreement between predicted values and observed values, and discrimination, the ability of a model to distinguish between patients with and without a specific outcome.

Calibration was quantified with the calibration slope[11,16]. The calibration slope was estimated as the regression coefficient of the linear predictor (lp). The lp was calculated by multiplying the predictor values of individual participants with the regression coefficients from the developed models: it is subsequently analysed as a single covariate or independent variable, with the outcome (either continuous outcome in a linear regression model or dichotomous outcome in a logistic model) as dependent variable. Models with perfect calibration should have a calibration slope equal to 1.

Discrimination was quantified with the area under the ROC curve or the concordance-statistic (c-statistic)[26]. All estimates of c-statistic compared predictions with the dichotomised outcome value. Predictions were either absolute values (continuous outcomes predicted with linear models) or risks (dichotomous outcomes predicted with logistic models). The absolute predicted values from a linear model were on the same scale as observed values (e.g. a predicted Hb value). ROC curves were constructed by estimating the sensitivities and specificities for all possible cutoff value of the predicted outcomes and subsequently plotting the sensitivities versus 1 minus specificities.

We also evaluated the performance of the models with R^2 (explained variance), an overall performance measure that incorporates both calibration and discrimination. We evaluated the explained variance also for dichotomised outcomes. R^2 was estimated as squared values of the Pearson correlation between the linear predictor from the models and the dichotomised outcome[27,28].

Internal and external validation

Internal validity of the prediction models was assessed with bootstrapping techniques. In every randomly drawn bootstrap sample, the entire modelling process including backward selection was repeated. The resulting model was then applied to the original sample to calculate the performance measures described in the previous section. The difference between the performance measures in the bootstrap sample and the performance measures in the original sample is a

measure for optimism of the prediction models. This process was repeated 100 times to obtain stable estimates of the optimism. The mean optimism in c-statistics and calibration slopes of the 100 bootstrapsamples were used to estimate the performance of models in new patients[16,29,30]. Prediction models derived with multivariable regression analysis are known for overestimated regression coefficients, which results in too extreme predictions when applied in new patients. We therefore used the calibration slope from the internal validation as a shrinkage factor[11,16].
External validity of the models with shrunken coefficients was assessed in patients from different general practices (depression) or participants from different blood collection centers (Hb deferral).

Results

Prediction models for depression

The highly skewed distribution of the PHQ-9 scores (Figure 1) resulted in non-normally distributed residuals from the full linear model. Linear models were therefore fitted with the Hubert-White 'sandwich' estimator to obtain robust standard errors[11]. Backward selection in the linear regression model eliminated three predictors: 'educational level', 'being single' and 'prescription of antidepressants in the past 12 months' (Table 1). Backward selection in the logistic regression model also eliminated 'being single', plus the 'number of life events', while 'educational level' and 'prescription of antidepressants' were retained. Note that the regression coefficients cannot directly be compared between linear and logistic regression models, because the methods use different scales. For example, women have a score on the PHQ-9 that is on average 0.6 higher than men. The odds for women to have a PHQ-9 score of 6 or higher is 1.3 (exp(0.3)) higher than for men.

Model performance

The discrimination of the models estimated in the same data that were used for development (apparent validity) was slightly higher for logistic models compared to the linear models (Table 2). The optimism in model performance was lower for the linear models. Calibration slopes at internal validation were closer to 1 for the linear models, e.g. the calibration slope was 0.932 for the full linear model and 0.870 for the full logistic model. Further, the differences in c-statistics and R^2 between apparent and internal validation were smaller for the linear models. The performance in the external validation samples were similar for linear and logistic models, with even a slightly better performance for the linear models.

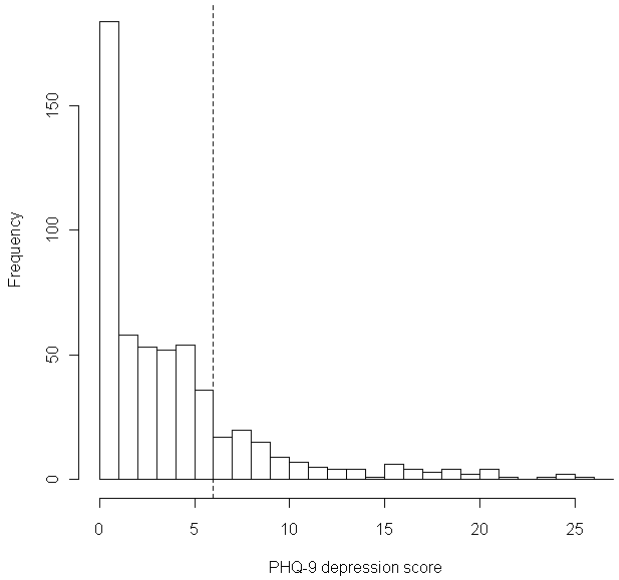


Figure 1. Distribution of PHQ-9 depression score
Legend: Distribution of PHQ-9 depression score. The dashed line indicates the cutoff level of 6.

Table 1. Regression coefficients (standard error) in clinically relevant depression models

	Linear regression*		Logistic regression*	
	Full model	Reduced model**	Full model	Reduced model**
Female gender	0.63 (0.32)	0.62 (0.32)	0.35 (0.26)	0.32 (0.26)
Age, per year	-0.02 (0.01)	-0.03 (0.01)	-0.02 (0.01)	-0.02 (0.01)
Educational level (None/primary only)	0.31 (0.45)	-	0.38 (0.28)	0.40 (0.27)
Being single	0.29 (0.48)	-	0.25 (0.28)	-
Number of presented complaints				
2	0.42 (0.53)	0.45 (0.52)	0.11 (0.32)	0.14 (0.32)
≥3	2.45 (0.92)	2.57 (0.87)	1.57 (0.49)	1.58 (0.48)
Non-somatic diagnosis/complaint at inclusion consult	0.45 (0.96)	-	0.68 (0.44)	0.73 (0.43)
Number of consults in the past 12 months	0.07 (0.03)	0.08 (0.03)	0.05 (0.02)	0.05 (0.02)
Received depression codes in past 12 months	3.96 (1.19)	4.37 (1.17)	1.26 (0.51)	1.29 (0.50)
Prescription of antidepressants in the past 12 months	0.77 (0.88)	-	0.60 (0.41)	0.60 (0.40)
Number of life events in the past 6 months				
1	0.16 (0.36)	0.18 (0.36)	-0.01 (0.30)	-
2	0.36 (0.46)	0.38 (0.46)	0.09 (0.33)	-
≥ 3	2.76 (0.70)	2.86 (0.71)	0.42 (0.33)	-
Any depressed feelings, lifetime	0.90 (0.51)	0.97 (0.52)	0.76 (0.33)	0.80 (0.33)
Any loss of interest, lifetime	1.24 (0.58)	1.22 (0.58)	0.79 (0.31)	0.77 (0.31)
Intercept	3.29	3.03	-2.77	-2.54

Note: Coefficients were shrunken, see methods section for details
- Not selected in the model
* Linear regression models predict the continuous PHQ-9 score; logistic regression models predict the risk that PHQ-9 score is 6 or higher.
** Model after backward selection with AIC (i.e. $p \leq 0.157$ for predictors with one regression coefficient)

Table 2. Performance of the models for clinically relevant depression

	Calibration Slope		Concordance-statistic*			R^{2}*		
	Internal validation	External validation[1]	Apparent	Internal validation	External validation	Apparent	Internal validation	External validation
Full model								
Linear	0.932	1.020	0.808	0.791	0.795	0.253	0.225	0.233
Logistic	0.870	0.941	0.824	0.799	0.786	0.273	0.232	0.221
Reduced model								
Linear model	0.929	0.992	0.799	0.781	0.782	0.243	0.214	0.215
Logistic model	0.863	0.887	0.820	0.794	0.768	0.266	0.222	0.198

[1] The calibration slope from internal validation was used to shrink regression coefficients before calculating the calibration slope in the external validation data
* For all models, the dichotomous outcome PHQ-9 ≥ 6 (y/n) was compared with the linear predictor

R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 R20 R21 R22 R23 R24 R25 R26 R27 R28 R29 R30 R31 R32 R33 R34 R35 R36 R37 R38 R39

Prediction models for Hb deferral

Standard linear regression analysis could be used for the analysis of continuous Hb values, because the values of Hb were normally distributed (Figure 2). Residuals from the linear model were also normally distributed. In this second example, prediction models were developed to detect low values of Hb and the regression coefficients of the logistic models thus have opposite signs compared with the linear models. In the linear models, higher predicted values are related to higher observed values, whereas higher predicted values (i.e. a higher predicted risk) are related to low values of Hb in the logistic model. Backward selection in the full linear model eliminated only 'Plasma donation in the past 2 years' (Table 3). Backward elimination in the logistic model also eliminated 'body mass index' (BMI). Although retained in the linear model because of statistical significance, BMI did not have clinical relevant predictive strength. A donor with a BMI of 20 has a predicted Hb level that is 0.02 mmol/l lower than a donor with a BMI of 25. Apparent, internal and external validation showed very minor differences between linear and logistic models. All calibration slopes (from internal and external validation) had values of 0.967 to 0.996, all c-statistics had values of 0.834 and R^2 values varied slightly between 0.08 to 0.10.

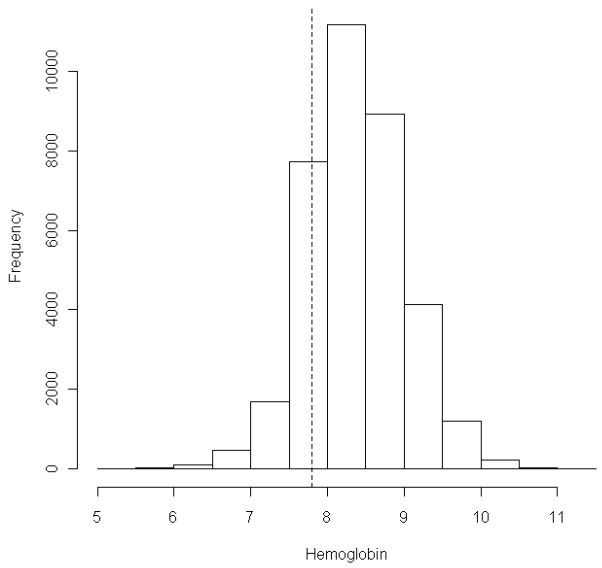


Figure 2. Distribution of hemoglobin
Legend: Distribution of hemoglobin. The dashed line indicates the cutoff level of 7.8.

Table 3. Regression coefficients (standard error) in Hb deferral models

	Linear regression*		Logistic regression*	
	Full model	**Reduced model****	**Full model**	**Reduced model****
Age, per 10 years	0.03 (0.02)	0.03 (0.02)	-.12 (0.02)	-.12 (0.02)
Seasonality[1]				
Spring	-.10 (0.01)	-.10 (0.01)	0.27 (0.06)	0.27 (0.06)
Summer	-.10 (0.01)	-.10 (0.01)	0.20 (0.06)	0.20 (0.06)
Autumn	-.002 (0.01)	-.002 (0.01)	-.20 (0.07)	-.20 (0.07)
Previous Hb level, mmol/l	0.69 (0.01)	0.69 (0.01)	-3.15 (0.07)	-3.15 (0.07)
Delta Hb, mmol/l	-.30 (0.01)	-.29 (0.01)	1.22 (0.04)	1.22 (0.04)
Time since previous visit, per 100 days	0.05 (0.002)	0.05(0.002)	-.32 (0.02)	-.32 (0.02)
Defferal at previous visit				
Because of low Hb	0.23 (0.01)	0.23 (0.01)	-1.05 (0.08)	-1.05 (0.08)
Because of other reasons	0.15 (0.01)	0.15 (0.01)	-1.11 (0.14)	-1.11 (0.14)
Number of whole blood donations in past 2 years	0.02 (0.003)	0.02 (0.003)	-.18 (0.02)	-.17 (0.02)
Plasma donation in the past 2 years, yes	0.06 (0.03)	-	-.24 (0.27)	-
BMI[2], kg/m^2	0.004 (0.001)	0.004 (0.001)	-.03 (0.01)	-
Blood volume, L	0.02 (0.007)	0.02 (0.007)	-.19 (0.06)	-.20 (0.05)
Intercept	2.14	2.14	26.07	26.07

[1]Winter is reference category
[2] BMI – Body Mass Index
- Not selected in the model

* Linear regression models predict the continuous Hb value; logistic regression models predict the risk that Hb is 7.8 or lower.
** Model after backward selection with AIC (i.e. $p \leq 0.157$)

Regression coefficients as estimated in the smaller subsample are shown in table 4. We observed some differences between these models and the models developed in the complete sample. For example, the regression coefficients in the linear model for 'Seasonality' were smaller, except the somewhat larger coefficient for the subcategory 'Autumn'. Backward selection in the linear model retained 6 predictors, while only three predictors were included in the logistic model.

Table 4. Regression coefficients (standard error) in Hb defferal models estimated in a small sample (EPV=5)

	Linear regression*		Logistic regression*	
	Full model	**Reduced model****	**Full model**	**Reduced model****
Age, per 10 years	0.02 (0.02)	0.03 (0.01)	-0.01 (0.01)	-
Seasonality[1]				
Spring	-0.05 (0.05)	-0.05 (0.05)	-0.03 (0.40)	-
Summer	-0.05 (0.05)	-0.05 (0.05)	-0.33 (0.40)	-
Autumn	0.07 (0.05)	0.07 (0.05)	-0.38 (0.41)	-
Previous Hb level, mmol/l	0.72 (0.04)	0.72 (0.04)	-2.20 (0.44)	-2.00 (0.34)
Delta Hb mmol/l	-0.31 (0.04)	-0.32 (0.03)	0.74 (0.30)	0.83 (0.29)
Time since previous visit, per 100 days	0.06 (0.02)	0.05 (0.01)	-0.37 (0.16)	-0.35 (0.02)
Defferal at previous visit				
Because of low Hb	0.41 (0.09)	0.40 (0.09)	-0.87 (0.57)	-
Because of other reasons	0.14 (0.09)	0.14 (0.09)	-0.81 (1.07)	-
Number of whole blood donations in past 2 years	0.01 (0.02)	-	-0.09 (0.12)	-
Plasma donation in the past 2 years, yes	0.25 (0.17)	-	-0.12 (1.14)	-
BMI[2], kg/m^2	0.003 (0.006)	-	0.05 (0.05)	-
Blood volume, L	0.01 (0.05)	-	-0.51 (0.39)	-
Intercept	1.96	2.07	18.34	14.86

Note: Coefficients were shrunken, see methods section for details
[1]Winter is reference category
[2] BMI – Body Mass Index
- Not selected in the model
* Linear regression models predict the continuous Hb value; logistic regression models predict the risk that Hb is 7.8 or lower.
** Model after backward selection with AIC (i.e. $p \leq 0.157$)

Model performance

In the small Hb development sample, discrimination was slightly higher for logistic models compared to the linear models (Table 5). The optimism in model performance was lower for the linear models. Calibration of the linear model was more stable in the external validation. The calibration slope of the full logistic model was even above 1 (1.048). Remarkably, the reduced linear and logistic models showed similar discriminative ability, while the linear model retained three extra predictors after selection.

Table 5. Performance of the model for Hb deferral – small sample (EPV 5)

	Calibration Slope		Concordance-statistic			R^2		
	Internal validation	External validation[1]	Apparent	Internal validation	External validation	Apparent	Internal validation	External validation
Full model								
Linear	0.977	0.974	0.801	0.794	0.821	0.077	0.073	0.073
Logistic	0.793	1.241	0.828	0.796	0.818	0.089	0.069	0.073
Reduced model								
Linear	0.978	0.974	0.801	0.794	0.819	0.077	0.073	0.072
Logistic	0.844	1.221	0.811	0.787	0.818	0.083	0.069	0.075

[1] The calibration slope from internal validation was used to shrink regression coefficients before calculating the calibration slope in the external validation data

R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 R20 R21 R22 R23 R24 R25 R26 R27 R28 R29 R30 R31 R32 R33 R34 R35 R36 R37 R38 R39

Discussion

In the large donor sample on Hb deferral, the differences in performance measures between the linear and the logistic models were negligible. We found no major advantages between predicting the original continuous outcome rather than the dichotomised versions. In small samples of two clinical examples, models predicting continuous outcome values showed better performance in terms of discrimination and calibration than models predicting dichotomous values. Prediction models for continuous outcomes showed less overfitting and retained more predictors in variable selection. In the depression example, when using PHQ-9 scores as a continuous outcome, the linear model showed a substantial better calibration in internal and external validation when compared to a logistic model using the dichotomised PHQ-9 score as outcome. In internal validation, discrimination of the linear model seemed lower than for the logistic model, in external validation, however, it was slightly better. In the small Hb deferral subsample, calibration of the linear models, when using the continuous Hb as outcome, was substantially better than the calibration of the logistic model and discrimination was slightly better.

In the depression example, we developed models with 11 candidate predictors (15 degrees of freedom) on 547 patients with 146 patients falling above the cuttoff score of the PHQ-9. In this example, the ratio of effective sample size to the number of degrees of freedom used in model development was 50 for the continuous outcome, whereas it was 10 for the dichotomised outcome. This considerable higher ratio for the continuous outcome resulted in a more stable (i.e. less overfitted) model with slightly better discrimination. In the Hb deferral example, we developed models with 10 candidate predictors (13 degrees of freedom) on 35687 participants with 3033 cases. In this example, the ratio of effective sample size to the number of degrees of freedom was 2745 for the continuous outcome compared to 233 for the dichotomised outcome. Hence, both samples were large in size and overfitting did not occur. Thus, when samples are very large, using a continuous outcome for development of prediction models is not necessarily better or worse than using a continuous outcome. In the small subsample of the Hb deferral sample a similar pattern as for the depression example was found. We developed models on 765 participants with 65 cases (ratio of effective sample size to the number of degrees of freedom was 59 and 5, respectively). The considerable higher ratio for the continuous outcome resulted again in less overfitting with slightly better discrimination. We expected a bigger difference in discrimination between the reduced linear and logistic models, since only three predictors were included in the logistic model and six predictors in the linear model. Previous Hb was a very strong predictor in all models and most likely explains the acceptable performance of the logistic model.

Several other studies noted various disadvantages of dichotomising continuous outcomes[6,9,12,14,15]. These studies, however, focussed on statistical analysis for etiological purposes, whereas this study evaluated the accuracy of prediction models intended for to support decision making in clinical

practice. For example, the prediction of depression models could be used to help treatment decisions, a decision that is dichotomous or categorical in nature. Other type of studies, such as clinical trials, may very well benefit from the increased effective sample size of continuous outcomes similar to the analysis of ordinal outcomes rather than dichotomous outcomes[31].

This study has some limitations. First, we presented the results of two case studies. As such, some findings may be specific, although realistic, for the examples presented here. Second, the choice of candidate predictors, derived from previous reports[17,25], that were included in the models may have influence the results. The inclusion of known predictors may have diluted differences in the performance between linear and logistic models. If more noise predictors were studied, larger differences between linear and logistic models may have been found. Third, we did not include models that consider the continuous outcomes together with the cutoff value for binary decisions, as proposed by Suissa et. al.[32]. This analysis was not included here, as a method to fit this model was not readily available. Further, for outcome variables that are not normally distributed, also other types of regression (e.g. regression with gamma distributions) may be preferable. Fourth, we focussed on statistical performance rather than clinical usefulness. We consider good statistical performance as a first important aspect of prediction models. Next clinical usefulness needs to be studied.

Conclusion

When prediction models are developed in sufficiently large effective samples, dichotomisation of continuous outcomes had no serious advantages or disadvantages. In smaller samples, the analysis of continuous outcomes is recommended to prevent development of too optimistic and overfitted prediction models.

References

(1) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338: b375.

(2) Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837-1847.

(3) Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS (2009) Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation 119: 3078-3084.

(4) Oudega R, Moons KG, Hoes AW (2005) Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 94: 200-205.

(5) Wells PS, Owen C, Doucette S, Fergusson D, Tran H (2006) Does this patient have deep vein thrombosis? JAMA 295: 199-207.

(6) Fedorov V, Mannino F, Zhang R (2009) Consequences of dichotomization. Pharm Stat 8: 50-61.

(7) Maxwell SE, Delaney HD (1993) Bivariate median-splits and spurious statistical significance. Psychological Bulletin 113: 181-190.

(8) Weinberg CR (1995) How bad is categorization? Epidemiology 6: 345-347.

(9) Breitling LP, Brenner H (2010) Odd odds interactions introduced through dichotomisation of continuous outcomes. J Epidemiol Community Health 64: 300-303.

(10) Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25: 127-141.

(11) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(12) Bhandari M, Lochner H, Tornetta P, III (2002) Effect of continuous versus dichotomous outcome variables on study power when sample sizes of orthopaedic randomized trials are small. Arch Orthop Trauma Surg 122: 96-98.

(13) Cuijpers P, Smit F, Hollon SD, Andersson G (2010) Continuous and dichotomous outcomes in studies of psychotherapy for adult depression: a meta-analytic comparison. J Affect Disord 126: 349-357.

(14) Streiner DL (2002) Breaking up is hard to do: the heartbreak of dichotomizing continuous data. Can J Psychiatry 47: 262-266.

(15) Suissa S (1991) Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol 44: 241-248.

(16) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(17) Zuithoff NP, Vergouwe Y, King M, Nazareth I, Hak E, Moons KG, Geerlings MI (2009) A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Fam Pract 26: 241-250.

(18) King M, Weich S, Torres F, Svab I, Maaroos H, Neeleman J, Xavier M, Morris R, Walker C, Bellon JA, Moreno B, Rotar D, Rifel J, Aluoja A, Kalda R, Geerlings MI, Carraca I, Caldas dA, Vincente B, Salidivio S, Rioseco P, Nazareth I (2006) Prediction of depression in European general practice attendees: the PREDICT study. BMC Public Health 6: 6.

(19) King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, Bellon-Saameno JA, Moreno B, Svab I, Rotar D, Rifel J, Maaroos HI, Aluoja A, Kalda R, Neeleman J, Geerlings MI, Xavier M, Carraca I, Goncalves-Pereira M, Vicente B, Saldivia S, Melipillan R, Torres-Gonzalez F, Nazareth I (2008) Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study. Arch Gen Psychiatry 65: 1368-1376.

(20) Kroenke K, Spitzer RL, Williams JB (2001) The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 16: 606-613.

(21) Spitzer RL, Kroenke K, Williams JB (1999) Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 282: 1737-1744.

(22) Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KG, Geerlings MI (2010) The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. BMC Fam Pract 11:98.: 98.

(23) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.
(24) Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453-473.
(25) Baart AM, De Kort WLAM, Atsma F, Moons KGM, Vergouwe Y (2011) Development and validation of a prediction model for low hemoglobin deferral in a large cohort of whole blood donors. Submitted .
(26) Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247: 2543-2546.
(27) DeMaris A (2002) Explained Variance in Logistic Regression: A Monte Carlo Study of Proposed Measures. Sociological Methods & Research 31: 27-74.
(28) Mittlbock M, Schemper M (1996) Explained variation for logistic regression. Stat Med 15: 1987-1997.
(29) Steyerberg EW, Eijkemans MJ, Habbema JD (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 52: 935-942.
(30) Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD (2001) Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 21: 45-56.
(31) Roozenbeek B, Lingsma HF, Perel P, Edwards P, Roberts I, Murray GD, Maas AI, Steyerberg EW (2011) The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. Crit Care 15: R127.
(32) Suissa S, Blais L (1995) Binary regression with continuous outcomes. Stat Med 14: 247-255.

Appendix 1. Characteristics of patients in the PREDICT-NL study

	Development (N=547)	Validation (N= 499)
Candidate predictors		
Female gender	63 (346)	66 (327)
Age, years[1]	44 (13)	45 (12)
Educational level (None/primary only)	22 (120)	19 (95)
Being single	21 (117)	24 (122)
Number of presented complaints		
1	81 (441)	77 (383)
2	14 (77)	17 (86)
≥3	5 (29)	6 (30)
Non-somatic diagnosis/complaint at inclusion consult	6 (35)	10 (52)
Number of consults in the past 12 months[2]	8 (4-14)	9 (5-15)
Received depression codes in past 12 months	6 (33)	5 (25)
Prescription of antidepressants in the past 12 months	9 (48)	8 (42)
Life events		
1	27 (145)	28 (141)
2	17 (93)	19 (97)
≥3	15 (81)	17 (83)
Any depressed feelings, lifetime	49 (267)	49 (243)
Any loss of interest, lifetime	39 (216)	40 (202)
Outcome		
PHQ-9 depression score[2]	3 (1-6)	3 (1-7)
PHQ-9 ≥ 6	27 (146)	30 (151)

Characteristics are reported als % (n), except where noted otherwise
[1] Mean (SD)
[2] Median (interquartile range)

Appendix 2. Characteristics of patients in the Hb defferal study

	Development (N=35687)	Validation (N= 29169)
Candidate predictors		
Age, years	45 (13)	45 (13)
Seasonality[1]		
Winter	19 (6805)	18 (5208)
Spring	26 (9119)	27 (7783)
Summer	30 (10702)	29 (8575)
Autumn	25 (9061)	26 (7603)
Previous Hb level, mmol/l	8.5 (0.6)	8.5 (0.6)
Delta Hb, mmol/l	-.03 (0.61)	-.02 (0.60)
Time since previous visit, months[2]	5 (4–8)	5 (4-8)
Defferal at previous visit[1]		
Low Hb	7 (2422)	6 (1813)
Other reasons	5 (1715)	4 (1178)
Number of whole blood donations in past 2 years[2]	3 (2-4)	3 (2-4)
Plasma donation in the past 2 years, yes[1]	1 (246)	0.4 (121)
BMI[3], kg/m^2	25 (4)	25 (4)
Blood volume, L	4.3 (0.5)	4.3 (0.5)
Outcome		
Hemoglobine, mean (SD)	8.4 (0.6)	8.5 (0.6)
Hemoglobine ≤ 7.8[1]	8 (3033)	7 (2034)

Characteristics are reported als mean (SD), except where noted otherwise
[1] % (n)
[2] median (interquartile range)
[3] BMI – Body Mass Index

10

General Discussion

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Concluding remarks

In the studies presented in part I, chapters 2, 3 and 4, we empirically examined the prediction of three very different outcomes in three different patient populations: presence of major depressive disorder in primary care patients, future occurrence of postherpetic neuralgia in patients diagnosed with herpes zoster infection and future occurrence of complications in patients who received a first pacemaker implantation.

At the start, little was known in the existing literature on potential predictors for developing postherpetic neuralgia. We therefore intially started with a large number of candidate predictors although the effective sample size of our dataset was small[1–4]. In the pacemaker implantation application, we observed limited discriminative ability of the candidate predictors, where also the effective sample size did not allow further evaluation of more predictors. Consequently, in both (prognostic) prediction studies we could only identify which were the most important, independently contributing predictors rather than set forward to develop a new prediction model for use in clinical practice. Such a model would be overfitted and too optimistic. For both examples we recommend that new studies are required to formally develop and (internally) validate these models, starting with the most promising predictors as found in our studies.

In the diagnostic example on the detection of major depressive disorder (chapter 4), we could develop and (internally) validate a formal diagnostic score for clinical use in primary care, because we were able to preselect a set of candidate predictors from the published literature and had ample number of outcomes. The empirical data at hand were thus mainly used to estimate the predictor effects of these candidate predictors, and not necessarily to select predictors from a large set of candidate predictors. We used multiple imputation techniques to deal with selectively missing values, which further allowed us to use all observed information for the model development[5–8]. In this way, the used data set was larger and more valid than when a complete case analysis had been performed.

In part II, we found that several key aspects of the methodology for clinical prediction research, which were also encountered in the empirical studies in Part I, are often not reported in sufficient detail. This is the main conclusion of our systematic review in chapter 6 and 7. This included details related to effective sample size, the number of degrees of freedom used in the development process, and how much missing data was present and how this was handled. Given this limited reporting, we could not judge if advanced statistical methods such as bootstrapping[9–12] should have been used in the reviewed studies. Internal validation[9,13], a strategy to assess overfitting and stability of developed models, was almost never reported. Even though missing values were often reported in the majority of studies, we also observed a high number of studies without a clear description of the number of missing values. Despite many methodological recommendations [4], the use of advocated methods, such as multiple imputation[5,7], were reported in only very

few studies. Most studies instead relied on the analysis of subjects with complete information, notwithstanding its potential for yielding biased results[14]. Another strategy to formally test the accuracy of prediction models, is by assessing the performance in new independent patients in so-called external validation samples[12,15–17]. Unfortunately, external validation of clinical prediction models was reported in very few studies in our review.

Methodological recommendations for the development of prediction models thus seem difficult to implement specifically in studies where the number of candidate predictors is relatively high in relation to the number of outcomes. We therefore further explored in Part II alternative statistical tools for model development in small samples. In chapter 7, we assessed the value of principal components analysis (PCA) for the reduction of the number of candidate predictors when developing a prediction model, either diagnostic or prognostic, in such contexts[18]. We found that this strategy can be valuable, as it reduces the number of coefficients (degrees of freedom) to be estimated in the analysis phase of the development of a prediction model, while retaining most of the information of the underlying predictors. In the presence of one strong predictor, LASSO (Least Absolute Shrinkage and Selection Operator)[19,20] or Ridge estimation[21], however, are better alternatives than PCA.

In chapter 8, we considered the development of prediction models for outcomes measured on a continuous scale, because the effective sample size for the analysis of a continuous outcome is larger than the effective sample size for the dichotomized version of such outcome[9,13,22]. Indeed the linear models predicting a continuous outcome performed better in new patients than the logistic models predicting a dichotomized version of such outcome if the development sample contained a limited number of subjects with the event after dichotomisation. The increase in effective sample size for a continuous outcome also allows researchers to evaluate more predictors and to study non-linearity and possible interaction. If an outcome is in essence continuous, we could more often consider to develop a prediction model for such continuous outcome and dichtomise afterwards. A disadvantage may be that the predicted outcome is also on a continuous scale (e.g. the predicted hemoglobin value). To classify patients correctly as having a low Hb may require a different cutoff value for the predicted outcome than the cut off value that is used to dichotomise the observed continuous value, due to the unexplained variance. This can be counterintuitive for the health care providers that will use the model.

Future research

We here focus on the perspectives of the methodological issues that were derived from part II of this thesis. The results of our systematic review clearly illustrate the need for guidelines on reporting, and possibly also for better conduct of prediction research, similar to existing guidelines for e.g. randomised trials, diagnostic studies and observational studies[23–26]. Guidelines for authors may subsequently promote the use of more advanced but proper methods in prediction research. Consistent reporting and conduct of prediction research also facilitates enhanced systematic reviews and perhaps even meta analyses of prediction studies. Additionally, consistent conduct and reporting of prediction studies facilitates meta analyses with individual patients data, an increasingly form of meta analyses in all types of medical research.

The application of principal components analysis in prediction research also need more research than the intial studies described in this thesis. PCA may be valuable for groups of predictors that, from a clinical or biological perspective, are related and could therefore reasonably be described by fewer components. We observed a relatively poor performance of this method in the presence of one strong predictor. The added value of other candidate predictors (or components from PCA) is then small and easily overestimated with PCA. LASSO and Ridge estimation are better alternatives in that situation. An important question is when a predictor is 'strong' and PCA should better not be considered. Additionally, we applied PCA to summarize predictive information from several predictors into fewer components. Another approach may be to use PCA to select predictors that best represent the components that explain a high percentage of the variance of the original variables[13], to achieve the best possible prediction with as few predictors as possible. Finally, the combination of LASSO and Ridge estimation, the so-called 'Naive Elastic Net'[27], has, to our knowlegde, not yet been evaluated in clinical prediction research at all. This combination may be valuable, as the LASSO tends to result in a too rigorous selection and Ridge estimation results in no selection at all.

The development of a prediction model with a continuous outcome, usually dichotomised for clinical use, should receive more attention in both the methodological research, as well as in empirical prediction research, as these methods make better use of the available information in the addressed outcomes. In chapter 8 we compared the statistical performance of linear and logistic models. Several issues, however, still need to be addressed. The first issue is whether correct decisions are indeed made based on the model's predicted continuous outcome value and how to determine relevant cut points to classify patients as having or not having the outcome at interest. Secondly, the use of the binary regression model for continuous outcomes as defined by Suissa et. al. is also minimally explored in empirical studies[28]. This is a promising technique since the distribution of the continuous outcome, the desired cut point and the location of the cut point within the distribution are all incorporated in the estimation method for the regression

coefficients. Furthermore, this method has been adapted to deviations from standard (e.g. normal) distributions[29]. Since simulation studies already showed in increase in statistical efficiency of this type of regression, it could be valuable for studies with a limited effective sample size[28,29]. Finally, problems may arise when a clinically relevant outcome is determined on more than one continuous outcome. For example, hypertension is usually diagnosed from systolic and diastolic blood pressure values. Statistical models suitable for the analyses of multiple outcomes, such as mixed models, may be used to analyse systolic and diastolic blood pressure simulataneously as outcomes, as well as their correlation[30]. The performance of this type of modelling and ways to convert the developed models into clinically useful prediction models needs to be evaluated.

Reference List

(1) Concato J, Peduzzi P, Holford TR, Feinstein AR (1995) Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 48: 1495-1501.

(2) Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48: 1503-1510.

(3) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379.

(4) Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

(5) Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59: 1087-1091.

(6) Gorelick MH (2006) Bias arising from missing data in predictive models. J Clin Epidemiol 59: 1115-1123.

(7) Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142: 1255-1264.

(8) Steyerberg EW, van Veen M (2007) Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 60: 979.

(9) Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

(10) Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54: 774-781.

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

(12) Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.

(13) Harrell, F. E. (5-6-2001) Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Springer Verlag.

(14) van der Heijden GJ, Donders AR, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59: 1102-1109.

(15) Toll DB, Janssen KJ, Vergouwe Y, Moons KG (2008) Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 61: 1085-1094.

(16) Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453-473.

(17) Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med 130: 515-524.

(18) Dunteman, G. H. (2011) Prinicpal Components Analysis. Newbury Park, Ca: Sage Publications Ltd.

(19) Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.

(20) Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society 58: 267-288.

(21) Le Cessie S, Van Houwelinge JC (1992) Ridge Estimators in Logistic Regression. Applied Statistics 41: 191-201.

(22) Fedorov V, Mannino F, Zhang R (2009) Consequences of dichotomization. Pharm Stat 8: 50-61.

(23) Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 340: c869.

(24) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 138: 40-44.

(25) Vandenbroucke JP, Von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med 4: e297.

(26) Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370: 1453-1457.

(27) Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. Biom J 52: 70-84.
(28) Suissa S, Blais L (1995) Binary regression with continuous outcomes. Stat Med 14: 247-255.
(29) Heritier S, Ronchetti E (2004) Robust binary regression with continuous outcomes. Canadian Journal of Statistics 32: 239-249.
(30) Snijders, S. and Bosker, R. (1999) Multilevel Analysis. London: Sage Publications Ltd.

Summary of this thesis

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Chapter 1 describes an overview of clinical prediction models from both clinical and methodological perspective. The empirical studies described in this thesis are divided in two parts. The first part describes the development of various prediction models.

In chapter 2, a clinical prediction rule with easily obtainable predictors for major depressive disorder in primary care patients is proposed. In the Dutch part of the PREDICT study (the PREDICT-NL study), 1046 subjects, aged 18 to 65 years, were included from seven large general practices in the center of the Netherlands. All subjects were recruited in the general practice waiting room, irrespective of their presenting complaint. Major depressive disorder according to DSM-IV criteria was assessed with the Composite International Diagnostic Interview (CIDI). Candidate predictors were gender, age, educational level, being single, number of presented complaints, presence of non-somatic complaints, whether a diagnosis was assigned, consultation rate in past twelve months, presentation of depressive complaints or prescription of antidepressants in past twelve months, number of life events in past six months, and any history of depression. The first multivariable logistic regression model including only predictors that require no confronting depression-related questions had a reasonable degree of discrimination (area under the ROC curve or c-statistic=0.71; 95% CI 0.67-0.76). Addition of three simple though more depression-related predictors, number of life events and history of depression, significantly increased the c-statistic to 0.80 (95% CI 0.76-0.83). After transforming this second model to an easily to use risk score, the lowest risk category (sum score < 5) showed a 1% risk of depression, which increased to 49% in the highest category (sum score≥ 30). The clinical prediction rule allows general practitioners to identify patients –irrespective of their complaints– in whom diagnostic workup for major depressive disorder is indicated.

In chapter 3, we present the reliability, construct validity and accuracy of the PHQ-9 and PHQ-2 to detect major depressive disorder in primary care. The PHQ-9 is especially attractive since the 9 diagnostic criteria for major depressive disorder are evaluated in a easy to use questionnaire. The PHQ-2 evaluates a subset of two of the most typical symptoms of major depressive disorder. The PHQ-9 showed a high degree of internal consistency (ICC=0.88) and test-retest reliability (correlation=0.94). With respect to construct validity, it showed a clear association with functional status measurements, sick days and number of consultations. The discriminative ability was good for the PHQ-9 (area under the ROC curve = 0.87, 95% CI: 0.84-0.90) and the PHQ-2 (ROC area = 0.83, 95% CI 0.80-0.87). Sensitivities at the recommended thresholds were 0.49 for the PHQ-9 at a score of 10 and 0.28 for a categorical algorithm. Adjustment of the threshold and the algorithm improved sensitivities to 0.82 and 0.84 respectively but the specificity decreased from 0.95 to 0.82 (threshold) and from 0.98 to 0.81 (algorithm). Similar results were found for the PHQ-2: the recommended threshold of 3 had a sensitivity of 0.42 and lowering the threshold resulted in an improved sensitivity of 0.81. The PHQ-9 is a reliable and valid measurement of DSM-IV major depressive disorder. Both the PHQ-9 and PHQ-2 had good discriminative ability. However, often recommended thresholds for the PHQ-9 and the PHQ-2 resulted in many undetected major depressive disorders.

In chapter 4, we studied which simple to measure factors are independent predictors of postherpetic neuralgia, and whether psychosocial and serological/virological parameters have additional predictive value. We included 598 elderly (>50 years) consecutive patients with acute herpes zoster (rash <7 days) below sixth cervical dermatome in a prospective cohort study in primary care. At baseline demographic, clinical (e.g. duration and severity of pain and rash), psychological (Pain Cognition List and Spielberger's Anxiety Inventory), serological and virological variables were measured. Blood tests were performed in a random subset of 218 patients. Primary outcome was significant pain after three months. The final prediction model obtained from multivariable logistic regression was (internally) validated using bootstrapping techniques, and adjusted for optimism. Forty-six (7.7%) patients developed postherpetic neuralgia. Independent predictors were age (OR =1.08 per year), acute pain severity (OR=1.02), presence of severe rash (OR=2.31), and rash duration before consultation (OR=0.78 per day): area under receiver-operating-characteristic curve was 0.77 (95% CI: 0.71-0.82). Of the five PCL scores, only factor V ('trust in healthcare') was an additional predictor (OR=1.01), though it increased the ROC area only 0.01 to 0.78. The Spielberger's anxiety scores and serological and virological variables were no additional predictors. Four simple predictors can help physicians to timely identify elderly herpes zoster patients at risk of postherpetic neuralgia.

In chapter 5 we assessed the incidence and determinants of short and long term complications after first pacemaker implantation for bradycardia in a prospective multicenter cohort study. The association of patient and implantation-procedure characteristics with the incidence of pacemaker complications were analysed using multivariable Cox regression analysis. In 23 Dutch PM centers, 1517 patients were followed for a mean of 5.8 years (SD 1.1), resulting in 8797 patient years. Within 2 months 188 (12.4%) patients developed a pacemaker complication. Male gender, age at implantation, body mass index, a history of cerebrovascular accident, congestive heart failure, use of anticoagulant drugs and passive atrial lead fixation were independent predictors for complications within 2 months, yielding a c-index of 0.62 (95% CI: 0.57-0.66). Thereafter, 140 patients (9.2%) experienced a complication, mostly lead related complications (n=84). Independent predictors for long term complications were age, body mass index, hypertension and a dual chamber device, yielding a c-index of 0.62 (95% CI: 0.57-0.67). A short term pacemaker complication was not predictive of future pacemaker complications. Complication incidence in modern pacing therapy is still substantial. Most complications occur early after pacemaker implantation. Although various characteristics are independent predictors for early and late complications, their ability to identify the patient at high risk is rather poor. This relatively high incidence of pacemaker complications and their poor prediction underscores the usefulness of current guidelines for regular follow-up of pacemaker patients.

The second part describes methodological issues in prediction modelling. In chapter 6 and 7, a systematic review of the methodology of prediction studies in the current literature is presented. In chapter 6, we investigated the reporting of aims, designs, participant selection, outcomes

and predictors, and statistical power of published prediction studies. We focused both on studies aimed at finding independent predictors of a particular outcome, and studies on the development or validation of prediction models. We did a full hand search to select all prediction studies published in 2008 in six general high impact journals. All types of multivariable prediction studies were included. Studies investigating causality were excluded. We developed an item list to score the study quality, based on recent recommendations for prediction research. We retrieved 71 papers for full text review, describing 135 diagnostic or prognostic prediction models. Only six studies addressed external validation of a previously developed model, and three the quantification of a model's impact on patient outcome. Eleven studies developed a new model; the remaining 51 were predictor finding studies. Study design was unclear in 16%, prospective cohort was used in 60%, retrospective cohort in 16% and case-control in 8%. Description and definition of participants, predictors and outcome was generally good. However, outcome measurement was blinded for predictor information in only 17%. Continuous predictors were dichotomized in 32% of studies. The number of events per variable (EPV) was undeterminable for 67%; of the remainder 53% had fewer than 10 EPV. We concluded that the vast majority of prediction studies include predictor finding studies, and only a few study external validity or impact of a prediction model. Despite various methodological recommendations for prediction research, the majority of prediction studies in high impact journals still do not follow these guidelines, limiting their reliability and applicability.

In chapter 7, we investigated the reporting of statistical methods, predictive performance and validation techniques of the published prediction studies. How candidate predictors were a priori selected was described in most studies (68%, n=44). A substantial number of studies relied on p-values of <0.05 for selection of predictors within univariable analysis. The presence of missing values was mentioned in 35% (n=24) of the studies, though most studies (89%, n=59) still conducted a so-called 'complete case analysis': only 5 studies used imputation methods to increase statistical efficiency and validity. Predictive model performance measures were reported in only 8 of the 68 modelling studies, with much variety in the type of measures. Calibration and discrimination was reported in 12% (n=8) and 27% (n=18) respectively. Multivariable effects from the final prediction model were usually presented, whereas the reporting of univariable effects was scarce. Considerable variation exists in the conduct and reporting of diagnostic and prognostic prediction studies, with respect to statistical methods, predictive performance and validation techniques. Poor statistical methods and poor reporting remain a threat to the transparency of prediction research and thus applicability of its findings in clinical practice.

In chapter 8, we illustrated prediction model development with principal components analysis (PCA), a method to combine predictors in components. We used two empirical examples, prediction of major depressive disorder and prediction of cardiovascular events. We developed models starting with a large number of candidate predictors. We compared four techniques of model development: 1. PCA; 2. backward selection; 3. Ridge regression; 4. LASSO (Least Absolute

Shrinkage and Selection Operator) estimation. The predictive performance of the models was evaluated in external data. In the major depressive disorder example, we observed the best calibration and discrimination for models developed with principal components analysis, Ridge and LASSO estimation. For the cardiovascular example, principal components analysis showed relative poor calibration compared to all other methods, due to one dominant predictor (age). Principal components analysis for the reduction of the number of predictors in prediction models showed similar performance as other advanced methods for model development. If a strong predictor dominates the predictive performance, principal components analysis may not be a good option for variable reduction.

In chapter 9, we investigated prediction of continuous outcomes. Continuous outcomes are frequently dichotomised and analyzed binary, as the clinical interest is often in discriminating patients that fall above or below a certain cut-off value. Equally as with continuous predictors, dichotomising outcomes may lead to loss in information and statistical efficiency (power). We developed prediction models for continuous outcomes with linear regression and then dichotomised the continuous outcomes and developed prediction models with logistic regression. We used the data of two clinical examples: prediction of clinically relevant depression and prediction of the presence of low hemoglobin level in blood donors. In the first example, prediction models were developed for clinically relevant depression defined as a high score on a depression scale (PHQ-9) in 547 primary care patients (27% high PHQ-9 score). In the second example, prediction models were developed for hemoglobin deferral (Hb deferral) defined as a low hemoglobin level in 35687 whole blood donors (8% low hemoglobin levels). In each example we compared the predictive performance of models developed with a continuous outcome with models developed for a dichotomised outcome. Internal validation (bootstrapping) was used to estimate optimism. Models were also externally validated. The performance was quantified with the *c*-statistic (or area under the ROC curve) and the calibration slope. Linear regression models for the prediction of high PHQ-9 scores showed less optimism in model performance than logistic regression (calibration slopes at internal validation 0.93 and 0.87, respectively). Discriminative ability of the linear model was slightly better (0.80 versus 0.79). Linear and logistic regression models showed similar performance for the prediction of low Hb in internal and external validation. When we decreased the effective sample size, linear regression models showed again less optimism and better external validity than logistic regression models. When prediction models are developed in sufficiently large samples, we found no major advantages between predicting the original continuous outcome rather than the dichotomised versions. In samples with few events, the use of linear regression analyses to predict the original continuous outcome may be recommended to prevent the development of too optimistic prediction models.

Samenvatting

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Het eerste hoofdstuk van dit proefschrift geeft een kort overzicht van predictiemodellen, beschouwd vanuit klinisch en methodologisch perspectief. De empirische studies in dit proefschrift zijn verdeeld in twee afzonderlijke delen. Het eerste deel, van hoofdstuk 2 tot hoofdstuk 5, beschrijft de ontwikkeling van een aantal predictiemodellen. Het tweede deel, van hoofdstuk 6 tot 8, gaat over methodologische aspecten van predictieonderzoek.

Deel een beschrijft de ontwikkeling van enkele klinische predictiemodellen. In Hoofdstuk 2 wordt een klinisch predictiemodel gepresenteerd met eenvoudige predictoren voor de diagnose 'depressieve stoornis'. Dit model werd ontwikkeld in het Nederlandse deel van de PREDICT studie (de PREDICT-NL studie). Deze studie betrof 1046 respondenten van 18 tot en met 65 jaar oud, die werden gerekruteerd met medewerking van zeven grote huisartspraktijken in Nederland. Alle respondenten werden benaderd in de wachtkamer van de huisartspraktijk, ongeacht hun reden voor consultatie. De diagnose 'depressieve stoornis' werd vastgesteld volgens de criteria van de DSM-IV met behulp van de *Composite International Diagnostic Interview* (CIDI). Kandidaat voorspellers die werden betrokken in het onderzoek waren: geslacht; leeftijd; opleidingsniveau; alleenstaand zijn; aantal klachten bij consultatie; presentatie van niet-somatische klachten bij consult; wel of geen diagnose na consult; aantal consulten in het jaar voorafgaand aan de consultatie; depressieve klachten in het jaar voorafgaand aan de consultatie; voorschrijven van medicatie voor depressieve klachten in het jaar voorafgaand aan de consultatie; aantal ingrijpende levensgebeurtenissen in het afgelopen jaar; en een eerdere diagnose van depressie in de medische voorgeschiedenis. Het eerste multivariabele logistische regressiemodel omvatte alleen die predictoren waarvoor geen vragen hoefden te worden gesteld die waren gerelateerd aan depressie. Dit model had een redelijk discriminerend vermogen (oppervlakte onder de ROC-curve, of *c-statistic*=0,71, 95% BI 0,67-0,76). Het toevoegen van drie depressiegerelateerde predictoren, ingrijpende levensgebeurtenissen en depressie in de voorgeschiedenis (twee vragen), leidde tot een significante verbetering van de *c-statistic* naar 0,80 (95% BI: 0,76-0,83). Dit tweede model werd omgezet in een simpel, eenvoudig te gebruiken scoresysteem. Respondenten in de categorie met de laagste score (somscore van vijf of minder) hadden een risico op depressie van 1%. In de hoogste categorie (somscore van 30 of meer) was dit opgelopen tot 49%. Deze predictieregel kan door huisartsen gebruikt worden om patiënten te selecteren voor wie verdere diagnostiek voor een depressieve stoornis geïndiceerd is.

Hoofdstuk 3 beschrijft de betrouwbaarheid, constructvaliditeit en accuratesse van de PHQ-9 en PHQ-2 voor van depressieve stoornissen bij huisartspatiënten. De PHQ-9 is een voor detectie aantrekkelijke en eenvoudig te gebruiken vragenlijst, die gebaseerd is op de negen criteria voor een depressieve episode. De PHQ-2 is een verkorte versie van de PHQ-9, waarin alleen naar de twee kenmerkendste criteria voor een depressieve stoornis worden gevraagd. Voor de PHQ-9 vonden wij een sterke interne consistentie (ICC=0,88) en hoge test-hertest correlatie van 0,94. We vonden een duidelijk verband met scores voor kwaliteit van leven, aantal ziektedagen en aantal consultaties bij de huisarts. Het discriminerend vermogen, uitgedrukt als de oppervlakte onder de

ROC-curve, was 0,87 (95% BI: 0,84-0,90) voor de PHQ-9 en 0,83 (95% BI: 0,80-0,87) voor de PHQ-2. De sensitiviteit voor de bestaande afkappunten van de PHQ-9 was 0,49 (score van tien of meer) en 0,28 voor het categorisch algoritme. Aanpassing van het afkappunt van tien en het algoritme verbeterde de sensitiviteit naar respectievelijk 0,82 en 0,84. Echter, de specificiteit zakte hierbij van 0,95 naar 0,82 (afkappunt tien) en 0,98 naar 0,81 (algoritme). De resultaten voor de PHQ-2 waren sterk vergelijkbaar: het aanbevolen afkappunt van drie had een sensitiviteit van 0,42 en het verlagen van het afkappunt verbeterde de sensitiviteit naar 0,81. De PHQ-9 is een betrouwbaar en valide instrument voor de DSM-IV-diagnose 'depressieve stoornis'. Zowel de PHQ-9 als de PHQ-2 hadden een goed discriminerend vermogen. Echter, de meest gebruikte afkappunten leidden tot grote aantallen niet gedetecteerde depressieve stoornissen.

In Hoofdstuk 4 onderzochten wij welke simpele predictoren gebruikt kunnen worden voor het voorspellen van postherpetische neuralgie. Daarnaast keken wij of psychosociale en serologische/virologische parameters toegevoegde waarde hebben voor het voorspellen van deze complicatie. Onze analyse omvatte 598 patiënten van 50 jaar of ouder met acute herpes zoster (minder dan zeven dagen uitslag) onder het zesde cervicale dermatoom in een prospectieve cohortstudie. Bij inclusie werden naast demografische en klinische (bv. duur en ernst van uitslag en pijn) gegevens ook de psychologische, serologische en virologische metingen geregistreerd. Bij 218 random geselecteerde patiënten werden tests gedaan op bloedmonsters. De primaire uitkomst was ernstige pijn na drie maanden. Het uiteindelijke predictiemodel, ontwikkeld met logistische regressie, werd intern gevalideerd met *bootstrapping*-technieken en gecorrigeerd voor optimisme. Postherpetische neuralgie deed zich voor bij 46 (7,7%) patiënten. Leeftijd, ernst van pijn bij opname in de studie, ernstige uitslag en het aantal dagen uitslag voor opname in het onderzoek waren onafhankelijke predictoren. Dit model had een oppervlakte onder de ROC-curve van 0,77 (95% BI: 0,71-0,82). 'Vertrouwen in de gezondheidszorg' was de enige van de psychologische factoren met voorspellende waarde, al was de toename in oppervlakte onder de ROC-curve slechts 0,01. Serologische en virologische voorspellers hadden geen toegevoegde waarde in het predictiemodel.

Hoofdstuk 5 onderzoekt de incidentie en determinanten van korte- en langetermijn complicaties na eerste pacemakerimplantatie voor bradycardie in een prospectieve cohortstudie in meerdere centra. Het verband tussen patiëntkenmerken en karakteristieken van de implantatieprocedure enerzijds en de incidentie van complicaties anderzijds werd geanalyseerd met multivariabele Cox-regressieanalyse. In 23 Nederlandse pacemakercentra werden 1517 patiënten gevolgd gedurende gemiddeld 5,8 jaar (SD 1.1). Honderdachtentachtig patiënten (12,4%) kregen in de eerste twee maanden een complicatie. Geslacht, leeftijd bij implantatie, Quételetindex, een eerder cerebrovasculair accident, hartfalen, antistollings medicatie en een passieve atriale draad waren predictoren voor complicaties binnen twee maanden. Dit model had een *c-statistic* van 0,62 (95% BI: 0,57-0,66). Daarna ontwikkelden 140 patiënten (9,2%) een complicatie, meestal een lead gerelateerde complicatie (n=84). Leeftijd, Quételetindex, hypertensie en een biventriculaire

pacemaker waren predictoren voor deze complicaties. Dit model had een *c-statistic* van 0,62 (95% BI: 0,57-0,67). Kortetermijn complicaties hadden geen voorspellende waarde voor langetermijn complicaties. Pacemaker implantatie leidt tegenwoordig nog steeds tot substantiële aantallen complicaties. De meeste complicaties deden zich kort na implantatie voor. We vonden een aantal predictoren van korte- en langetermijn complicaties: de voorspellende waarde (i.h.b. het discriminerend vermogen) was echter onvoldoende. De relatief hoge incidentie van complicaties en de ontoereikende voorspellende waarde van de predictoren benadrukt het belang van de huidige richtlijnen voor regelmatige controle van patiënten met een pacemaker.

Het tweede deel van dit proefschrift gaat in op methodologische aspecten van predictieonderzoek. Hoofdstuk 6 en 7 presenteren een systematische kritische bespreking van predictieonderzoek in recente medische literatuur. In Hoofdstuk 6 bespreken we de rapportage van het doel, de opzet, de selectie van deelnemers, de uitkomst, de voorspellers en de statistische power van gepubliceerde predictiestudies. We beperkten ons tot in 2008 gepubliceerde predictiestudies gericht op het identificeren van onafhankelijke predictoren en studies gericht op het ontwikkelen of valideren van predictiemodellen. Deze werden met de hand geselecteerd uit zes algemene medisch-wetenschappelijke tijdschriften met een hoge impactfactor. Alle typen multivariabele predictiestudies werden geïncludeerd. Artikelen waarin causale relaties werden bestudeerd werden geëxcludeerd. We ontwikkelden een itemlijst voor het scoren van de kwaliteit van de studies op basis van recente aanbevelingen voor predictieonderzoek in de methodologische literatuur. We vonden 71 studies waarin in totaal 135 diagnostische of prognostische modellen werden beschreven. Slechts zes van deze studies werd de externe validiteit van een eerder ontwikkeld predictiemodel beschreven. In drie studies werd de impact van predictiemodellen op patiëntuitkomsten beoordeeld. Elf studies ontwikkelden een nieuw model. De overige 51 beschreven onderzoeken waarin gezocht werd naar voorspellers (de zgn. 'predictor finding studies'). In 16% van de artikelen was het onderzoeksontwerp van de studie onduidelijk beschreven. Zestig procent van de studies had een prospectief ontwerp, 16% een retrospectief ontwerp en 8% een case-control design. De selectie van deelnemers, predictoren en uitkomst waren in het algemeen goed beschreven. Echter, in slechts 17%, was de predictor geblindeerd voor het bepalen van de uitkomst. Continue predictoren waren gedichotomiseerd in 32% van de studies. Het aantal *events* per variabele (EPV) was in 67% van de studies niet te bepalen; 53% van de overige studies had een EPV van minder dan tien. We concluderen dat het grootste deel van de predictiestudies zoekt naar voorspellers van een specifieke uitkomst. Een klein aantal studies kijkt naar de externe validatie of de impact van een predictiemodel. Aanbevelingen uit de methodologische literatuur worden door een groot deel van de predictiestudies niet opgevolgd, waardoor de betrouwbaarheid en toepasbaarheid van deze studies onvoldoende duidelijk is.

In Hoofdstuk 7 onderzoeken we de rapportage van statistische methoden, statistische maten voor het voorspellend vermogen en de validatietechnieken van de gepubliceerde predictiestudies.

De meeste onderzoeken beschreven de selectie van kandidaat predictoren (68%, n=44). Een aanzienlijk aantal studies selecteerde vervolgens predictoren op basis van p-waarden van 0,05 of minder in univariabele analyses. Missende waarden werden gerapporteerd in 35% (n=24) van de artikelen. De meeste onderzoeken (89%, n=59) gebruikten voor de statistische analyse alleen de deelnemers voor wie volledige gegevens beschikbaar waren. Niet meer dan vijf studies gebruikten imputatiemethoden om de statistische efficiëntie en validiteit te verbeteren. Slechts acht van de 68 onderzoeken rapporteerden maten voor de voorspellende waarde van de predictiemodellen, met een grote variatie in het type maat dat gerapporteerd werd. Calibratie- en discriminatiematen werden in niet meer dan (respectievelijk) 12% (n=8) en 27% (n=18) van de studies gerapporteerd. De meeste predictiestudies presenteerden resultaten van het uiteindelijke ('finale') predictiemodel. Een klein aantal artikelen lieten ook univariabele resultaten zien. Wat betreft statistische methoden, het voorspellend vermogen en de validatietechnieken is er veel variatie in zowel de rapportage als in de manier waarop die worden getest. Tekortschietende statistische methodes en ontoereikende rapportage van deze technieken leiden tot een gebrek aan transparantie in predictieonderzoek en derhalve aan de toepasbaarheid van de resultaten in de klinische praktijk.

Hoofdstuk 8 illustreert de ontwikkeling van een predictiemodel met principale componenten analyse (PCA), een methode om predictoren te combineren in zogenaamde 'componenten'. We presenteerden twee voorbeelden: predictie van de diagnose 'depressieve stoornis' en predictie van cardiovasculaire *events*. We ontwikkelden modellen met een groot aantal predictoren. Hierbij vergeleken wij vier technieken voor het ontwikkelen van deze modellen: 1. PCA; 2. Standaard achterwaartse selectie ('backward selection'); 3. Ridge-regressie; 4. de LASSO ('Least Absolute Shrinkage and Selection Operator'). Het voorspellend vermogen van deze modellen werd geëvalueerd in externe data. In het depressie-voorbeeld vonden we het beste voorspellende vermogen bij gebruik van PCA, Ridge-regressie en de LASSO. In het cardiovasculaire voorbeeld gaf PCA een relatief slecht resultaat in vergelijking met alle andere methoden. Dit werd veroorzaakt door de zeer sterke rol van de predictor leeftijd. Modellen ontwikkeld met PCA hebben een vergelijkbaar voorspellend vermogen als andere geavanceerde methoden. Bij de ontwikkeling van modellen met één zeer sterke predictor, echter, is PCA een minder goede optie.
In Hoofdstuk 9 kijken we naar het voorspellen van continue uitkomsten. Continue uitkomsten worden vaak gedichotomiseerd en vervolgens geanalyseerd als een binaire uitkomst, omdat dit het beste aansluit bij toepassingen in de klinische praktijk. Dichotomiseren kan echter leiden tot verlies van informatie en statistische efficiëntie (power). We ontwikkelden predictiemodellen voor continue uitkomsten met lineaire regressie. Vervolgens werden deze uitkomsten gedichotomiseerd en ontwikkelde we modellen met logistische regressie. Hierbij gebruikten wij data van twee klinische voorbeelden: het voorspellen van klinisch relevante depressie en lage hemoglobinewaarden bij bloeddonoren. In het eerste voorbeeld ontwikkelden we een

predictiemodel voor depressie, gedefinieerd als een hoge score op een depressieschaal (de PHQ-9), met gegevens van 547 huisartspatiënten. In het tweede voorbeeld maakten we predictiemodellen voor een te laag hemoglobineniveau (minder dan 7,8 mmol/l) bij 35687 bloeddonoren. In beide voorbeelden vergeleken we het voorspellend vermogen van de modellen ontwikkeld met de continue uitkomst met de modellen ontwikkeld met de dichotome uitkomst. Interne validatie werd gebruikt om optimisme te schatten. De modellen werden ook extern gevalideerd. Het voorspellend vermogen werd bepaald met de zogenaamde *c-statistic* (ook: oppervlakte onder de ROC-curve) en met de calibratiecoëfficiënt (*calibration slope*). De lineaire regressiemodellen voor het voorspellen van hoge PHQ-9 scores lieten minder optimisme in voorspellend vermogen zien dan de logistische regressiemodellen (calibratiecoëfficiënten van 0,93 en 0,87 respectievelijk). Het discriminerend vermogen van het lineaire model was iets beter (0,80 vs. 0,79 respectievelijk). De lineaire en logistische modellen voor het voorspellen van hemoglobinewaarden waren nagenoeg gelijk bij interne en externe validatie. Bij de ontwikkeling van modellen in een kleine subset vonden we minder optimisme en een betere externe validiteit van het model gebaseerd op lineaire regressie. Als predictiemodellen worden ontwikkeld in datasets met grote aantallen deelnemers, biedt het analyseren van een continue uitkomst geen voordelen. Maar in data met kleine effectieve steekproefgrootte is het ontwikkelen van een predictiemodel met lineaire regressie aan te bevelen, om optimisme tegen te gaan.

Dankwoord

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

A long time ago in a galaxy far, far away, ben ik begonnen bij (toen nog) huisartsgeneeskunde. In alle eerlijkheid dacht ik dat ik solliciteerde op een tijdelijk baantje om wat data in te voeren. Het "baantje" groeide vrij snel uit tot het ondersteunen van (meestal) klinische onderzoekers bij data-analyse. Dit is langzaamaan doorgegroeid naar dit proefschrift. Dit woord van dank is voor een groot aantal mede Julianen, natuurlijk aan ieder die aan dit proefschrift heeft bijgedragen, maar ook aan een grote groep collega's met wie ik de afgelopen jaren heb mogen samenwerken.

Prof. dr. K.G.M. Moons, beste Carl, meer dan bij welke andere promovendus zou dit proefschrift er niet geweest zijn zonder jouw steun, jouw inzet, en jouw nimmer aflatende vertrouwen in mij. Daarnaast vind ik je een 'motivational genius': iemand die kan enthousiasmeren en kan motiveren, ook op momenten dat ik zelf vond dat ik met een project volledig klem zat. Dank voor je vertrouwen. *Meet me across the river(s)*, niet 'on the Jersey side', maar in het verre zuiden voor een avond van goede rock&roll.

Dr. Y. Vergouwe, beste Yvonne, ik heb ontzettend veel van jou geleerd over analyse, predictie, optimisme, R, performance en validiteit en subtiele (maar belangrijke) verschillen en overeenkomsten daartussen. Dat was goed voor mijn proefschrift, afgezien daarvan heb ik er ook veel aan gehad voor allerlei andere projecten. Heel veel dank voor je inzet, je geduld en je betrokkenheid.

Dr. M.I. Geerlings, beste Mirjam, ook zonder jou zou dit proefschrift er niet zijn geweest: je bent degene die PREDICT heeft gedragen, de studie die in een groot deel van dit proefschrift terugkomt. Je hebt me laten vaak zien hoe teksten voor artikelen korter, duidelijker en overzichtelijker gemaakt konden worden. Ik heb erg veel bewondering voor je gedrevenheid en je heldere inzicht en verstand. Ik vond het ook erg leuk om af en toe mee te kijken bij andere (deel)projecten, zowel binnen als buiten Predict. Heel veel dank voor je inzet, je geduld en betrokkenheid.

Geachte commissie, prof. dr. A.W. Hoes, prof. dr. N.J. de Wit, prof. dr. H.F.E. Smit, prof. dr. H.J.A. Hoijtink, dank voor de bereidheid om tijd vrij te maken voor het lezen en beoordelen van mijn manuscript. Beste Niek, ik ben vanaf het begin van mijn werk bij huisartsgeneeskunde bij jouw onderzoeksprojecten betrokken geweest: het is voor mij daarom extra leuk dat jij bereid was om in de commissie plaats te nemen. Beste Arno, ook met jou heb ik veel samengewerkt, je hebt bovendien mijn promotie mede geïnitieerd. Dank voor het vertrouwen en de goede samenwerking.

Bijna dr. W. Bouwmeester, beste Walter. Je bent een geweldige collega om een zware klus als een "systematic review" uit te voeren. De review was in eerste instantie jouw project: ik heb mede geprofiteerd van het voorwerk wat jij al gedaan had, en van je goedmoedigheid, je tolerantie voor

al mijn flauwe grappen, je doorzettingsvermogen en integriteit. Succes met jouw laatste loodjes en vooral: Dank! *Don't be a stranger....*

Heel veel dank aan de 'Predict' groep, de collega's die zich hebben ingespannen om de gegevens te verzamelen en de mede-onderzoekers die de Predict data verder geanalyseerd hebben: Manja van Wezep, Hanneke den Breeijen, Marjolein Kamphuis, Bauke Stegenga, Anne Grool en Thea Haks.

Many thanks to all co-authors: prof. dr. M. King, prof. dr. I. Nazareth, prof. dr. E. Hak, dr. S. Mallet, prof. dr. E.W. Steyerberg and prof. dr. D.G. Altman for their valuable contributions.

Prof. dr. Y vd Graaf, beste Yolanda, ik mocht voor één van methodologische studies gebruik maken van de Smart data. Dank aan jou en aan de Smart studiegroep.

Mireille Baart, ook hard op weg om dr. te worden en dr. W.L.A.M. de Kort. Ik wil jullie en de andere leden van de 'Donor groep' danken dat ik voor mijn laatste stuk gebruik kon maken van de data van jullie studie.

Tijdens mijn promotie heb ik meegewerkt aan de Followpace studie, waarvan ik één artikel mocht opnemen in mijn proefschrift. Followpace werd geïnitieerd door prof. dr. Van Hemel, het project werd mede 'getrokken' door Martijn van Eck en (nu) door Erik Udo.

Dr. A.R.T. Donders, beste Rogier, ik leerde je kennen als iemand bij wie ik af en toe terecht kon met echt ingewikkelde statistische vragen (die je altijd kon beantwoorden) en daarna als iemand die ik kon lastig vallen met R vragen en die kans ziet om op een congres in minder dan 5 minuten allerlei suggesties voor zowel mijn onderzoek als voor iemand anders doet. Eén daarvan is terecht gekomen in het PCA stuk, waar je nu mede auteur bent. Dank voor je bijdragen en enthousiasme.

De collega's van mijn 'oude baan' bij ooit het SVUH, later TEO, en nog later weer wat anders. Ondanks de soms woelige tijden vaak gezellig en saamhorig, ik mis jullie nog steeds! Elbrig Pasma, Paula Jobse, Koos Jansen, Anneke Kramer, Wilma Spinnewijn en Gerda Weeda, Peter Thys, Jaap Buijs, Henk Folkers en Lisa Tan dank voor de goede tijd! Herman Düsman is een bijzondere collega voor mij geweest: iemand die mij meer dan wie ook gestimuleerd heeft in systematisch en logisch nadenken voor wetenschappelijk onderzoek. Herman, dank voor je inzet, collegialiteit en vriendschap.

Ik heb lange tijd bij de afdeling datamanagement gewerkt. Ik heb in die tijd drie leidinggevenden gehad: Hanneke den Breeijen, Robert Veen en Frank Leus. Van ieder van jullie heb ik altijd de kans

gehad om leuke projecten te doen, vaak projecten die ik zelf aandroeg en/of graag wilde doen. Elk van jullie heeft veel vertrouwen gehad in mijn werkzaamheden en mij de mogelijkheden geboden om mezelf te ontwikkelen. Zonder die mogelijkheden zou dit proefschrift er niet geweest zijn. Naast leidinggevenden waren er natuurlijk de collega's en en ex-collega's: Ronald, Eefje, Marloes, Frank, Alexander, Bram, Jan-Willem, Janneke, Susan, Julia, Jildou en Marianne. Met een aantal van jullie heb ik ook een kamer gedeeld: Nicole, Yen, Joost, Diane, Lex, Lara dank voor de gezelligheid en lol op de werkkamers. Dank ook aan Rutger, oud kamergenoot en degene die vaak klussen waarnam in de tijd dat ik te druk met schrijven was. Vorig jaar verloren wij collega Bernard Slotboom, een erudiet en zeer prettig mens, iemand met het meest door ontwikkelde gevoel voor humor die ik ooit ontmoet heb.

Aan het begin van mij traject kwam ik terecht bij de 'predictieclub', later omgedoopt tot de 'methodologieclub'. Zonder twijfel de meest 'nerdy', oftewel de slimste onderzoeksclub van het Julius en een groep waar ik veel van geleerd heb. Veel dank aan leden en ex-leden: Rolf, Geert, Erik, Lidewij, Linda, Thomas, Stan, Kristel, Diane, Liselotte, Roelof, Susan, Sjoukje, Joris, Roelof, Wouter, Teus, Charlotte, Merel, Christa, Anne-Mette, Henrike, Sabrina, Suleyman, Janneke, Floriaan, Maarten, Anoukh en eenieder die ik vergeet, maar die zeker ook bij heeft gedragen. Hans Reitsma, dank voor jou gedegen introductie in meta-analyse en je hulp bij mijn eerste schreden op dit gebied.

Heel veel dank aan Elemi Breetvelt, Marco Boks en alle andere leden van de voormalige 'Neuro-epi' groep.

Eén van de eerste 'leermeesters' die ik heb gehad was Aart Lodder, die zijn degelijke en gedetailleerde kennis op het gebied van methoden en statistiek met mij deelde. Beste Aart, veel dank voor jouw inzet indertijd. Je hebt zonder dat wij dat beiden beseften een fundering gelegd voor een groot deel van mijn latere prestaties, niet in de laatste plaats voor dit proefschrift! Een andere belangrijke collega, meer nog een toeziend oog, op mijn werk was Carla Tims. Beste Carla, ik profiteer nog dagelijks van jouw feilloze oog voor precisie en detail, iets wat onontbeerlijk is in goed onderzoek.

Vanaf het begin ben ik betrokken geweest bij een groot aantal onderzoeksprojecten, vooral van de huisarts-onderzoekers. Dit waren (en zijn) projecten waar ik met veel plezier aan gewerkt heb: ik heb de huisarts-onderzoekers leren kennen als onderzoekers met veel gedrevenheid en (vooral) veel liefde voor het zowel de zorg voor hun patiënten als voor het onderzoek. Ik weet zeker dat ik projecten en onderzoekers ga vergeten, ook die dat zeker niet verdienen, het zijn in de loop der jaren veel projecten en veel onderzoekers geweest. Een poging, in een helemaal 'gerandomiseerde' volgorde: Frans Rutten, Kees Gorter, Otto Quartero, Roger Damoiseaux, Lex Goudswaard, Wim

Opstelten, Ted van Essen, Catherine Weijnen, Tjarda Scheltens, Rykel van Bruggen, Frank van Balen, Bert-Jan de Boer, Pieta Bruggink, Mattijs Numans, Rene Bijkerk, Marieke Dekker, Madeleine Bruins-Slot, Annemieke Akkermans, Huug van Duijn, Hugo Smeets, Paul Janssen, Roeland Geijer, Margit Vermeulen, Michelle Westerhuis, Martijn van Eck, Saskia van Vugt, Frits Cleveringa, Bas van Zaane, Nicky Peters, Ineke Welschen, Nadine Goessens, Tim Timmers, Irène Oudejans, Geert-Jan Geersing, Ewoud Schuit.

Dr. M.M. Kuyvenhoven, beste Marijke. Ik heb met jou letterlijk vanaf dag 1 onderzoeks projecten gedaan. Antibioticaprescripties, attitudes, vignetten, vragenlijstontwikkeling en meest recent een project over onderwijs. Jij bent veelzijdig in wat je onderzoekt, en stimuleerde mij om me te verdiepen in wat we nodig hadden aan statistische analyses voor al deze projecten. Naast veelzijdig ben je ook uitgesproken prettig, hartelijk en belangstellend: het was een groot genoegen om aan deze projecten te werken.

Prof. dr. G.E.H.M. Rutten, beste Guy, een substantieel deel van de onderzoeksprojecten die ik vanuit datamanagement heb gedaan werd uitgevoerd onder jouw leiding. Ik heb de afgelopen jaren met groot genoegen meegewerkt aan diverse projecten in de diabetologie.

Prof. dr. Th. J.M. Verheij, beste Theo. Jij bent één van mensen die ergens aan de wieg van dit proefschrift staat, toen je een aantal jaar geleden eens voorzichtig informeerde of ik belangstelling zou hebben om zelf te promoveren. Het resultaat ligt nu voor je. Ik ben je erg dankbaar voor zowel die aanzet als voor alle projecten die ik heb kunnen doen in het huisartsgeneeskundig onderzoek.

Heel veel dank aan Annina Koopmans, Marlies Blijleven, Cootje Kusters en alle andere collega's van het Julius.

Een groot aantal vrienden hebben de ins-en-outs van het promotieleven aanzienlijk dragelijker gemaakt, ondanks dat hen de laatste tijd behoorlijk verwaarloosd heb. Mirjam & Jochem en Aaron en Isis, Peter H., Ruud, Marlies, Oscar, Sandra, Hans, Frank, Kaspar, ik ben er minder vaak bij geweest dan ik had gewild. Desalniettemin waren ieder van jullie belangstellend naar mijn vorderingen. Remco en Tonny, ook jullie dank voor belangstelling, feestjes, etentjes, veel smakeloze maar erg leuke grappen en veel Bruce! Heel veel dank ook aan Henk Sluiters, een *late arrival* in mijn vriendenkring, een fijn mens en een goede 'aanwinst'.

Heel veel dank voor ontspanning en gezelligheid en 'oppasfeestjes' aan Wendelmoet & Robert. Ik ben af en aan door jullie achter de computer vandaan gehaald voor een etentje, een filmpje, een dagje uit, of gewoon een uurtje oppassen, wat ontzettend veel hielp om het vol te houden.

Berre en Froukje hebben mij er meermaals op gewezen dat lego, dino's, drakenridders, prinsessen en knutsels onvoorstelbaar veel leuker en belangrijker zijn dan artikelen en tabellen. Dank, jullie hebben groot gelijk!

Heel veel dank aan Giselinde voor nu meer dan 20 jaar trouwe vriendschap. We zijn na onze studietijd op heel verschillende plekken terecht gekomen, maar delen enthousiasme voor wetenschappelijk onderzoek. Ik vond het geweldig om je te zien oreren, ik ben trots en dankbaar jou tot mijn beste vrienden te kunnen rekenen! Dank ook aan de *late great* Gibraltar, een incidentele maar gezellige logé, het enige levende wezen dat ik ooit ontmoet heb dat nog langer op de bank kan hangen dan ikzelf. Hij wordt zeer gemist.

Van al mijn vrienden was Maarten degene die allerlei aspecten van mijn onderzoek het gemakkelijkst kon volgen. Veel dank, Maarten voor het meedenken, voor de gezelligheid, het regelmatige zondagse eten, voor je belangstelling, betrokkenheid, de optredens en voor de nu meer dan 20-jarige vriendschap. Ik ben blij en trots dat jij naast me staat als paranimf!

Pauline, ik ben ook blij en trots dat jij naast me staat als paranimf. Je hebt veel bijgedragen aan dit proefschrift, misschien niet direct aan de inhoud, maar wel met een filmpje, een avondje cultuur, simpelweg bijpraten, een dagje fietsen, wandelen of wat dan ook. Ik kan echt opleven van je gevoel voor humor en je berg flauwe, botte en vooral hele leuke grappen. Blijf ze maken. Ook met jou staat de teller op 20+, en ik ben nog lang niet uitgeteld.

Het zwaarste deel van mijn promotie was de tijd dat mijn beide ouders zijn overleden, mijn moeder in 2007 en mijn vader in 2008. Ik heb altijd gehoopt dat zij mijn promotie nog mee zouden kunnen maken, maar het heeft niet zo mogen zijn. Ik ben hen veel verschuldigd, ook en vooral het deel waarin zij mij stimuleerde om mijn eigen weg te vinden in mijn opleiding en later in de wetenschap.

Jarenlang door mij persoonlijk uitgevoerd onderzoek heeft uitgewezen dat er maar één echte 'oom Kees' bestaat, namelijk die van mij en mijn broers. Altijd belangstellend, zeer bescheiden en bovenal een authentiek vriendelijk en goed mens. Dank, oom Kees, voor je interesse en je medeleven. Lieve Deborah, Mireille, ik heb me bij elk van jullie altijd ontzettend welkom gevoeld. Dank voor jullie hartelijkheid en belangstelling. Ik had me geen betere en leukere schoonzussen kunnen wensen! Chantal en Stefan, ook voor jullie een dankjewel: Peet heeft het boek af, ik kan weer wat vaker langskomen. Lieve Quinten, welkom klein wonder, ik ben blij dat je er bent. Mijn grote broer Kees, en mijn grote broer(tje) Herman, dank voor jullie support, aandacht, hulp, enthousiasme, en niet in de laatste plaats voor *1001 night of good comradeship and rockin'.*

Curriculum Vitae

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Nicolaas Pieter Adriaan Zuithoff, Peter to most who know him, was born on june 22th 1967 in Rotterdam. He graduated in 1987 from the scholengemeenschap "Comenius". In 1989, he started his studies in psychology at the University of Utrecht, where he graduated in 1995. In 1997, he started working for the department of General Practice, part of the medical faculty of the University of Utrecht, which became part of the Julius Center for Health Sciences and Primary Care. As part of his work, he started working on his thesis in 2005 under supervision of prof. dr. K.G.M. Moons, dr. M.I. Geerlings and dr. Y. Vergouwe. His primary interests are applied statistic with a special interest in prediction models.