Improvements in clinical prediction research

Kristel Josephina Matthea Janssen

Improvements in clinical prediction research

Utrecht, Universiteit Utrecht, Faculteit Geneeskunde
Thesis, with a summary in Dutch
Proefschrift, met een samenvatting in het Nederlands

ISBN	978-90-393-4668-6
Author	Kristel J.M. Janssen
Lay-out	Roy Sanders
Cover design	Jacco van Muiswinkel, Websites-en-co
Print	Gildeprint BV, Enschede

We gratefully acknowledge the support by the Netherlands Organisation for Health Research and Development (ZonMw 016.046.360)

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

Improvements in clinical prediction research

Ontwikkelingen in klinisch predictieonderzoek
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de Rector Magnificus, Prof. Dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 6 december 2007 des ochtends te 10.30 uur

door

Kristel Josephina Matthea Janssen

geboren op 10 november 1975 te Eijsden

Promotores Prof. Dr. K.G.M. Moons
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht

Prof. Dr. D.E. Grobbee
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht

Co-promotor Dr. Y. Vergouwe
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht

Manuscripts based on the studies presented in this thesis

Chapter 2.1
K.J.M. Janssen, A.R.T. Donders, F.E. Jr. Harrell, Y. Vergouwe, Q. Chen, D.E. Grobbee, K.G.M. Moons.
Missing data in medical research: making up is better than giving up.
Submitted.

Chapter 2.2
K.J.M. Janssen, I. Siccama, Y. Vergouwe, M. Keijzer, D.E. Grobbee, K.G.M. Moons.
Derivation methods for clinical prediction models: a comparison between conventional logistic regression, penalised maximum likelihood estimation and genetic programming.
Submitted.

Chapter 3.1
D.B. Toll, K.J.M. Janssen, Y. Vergouwe, K.G.M. Moons.
Validation, Updating and Impact of clinical prediction rules: a review.
Accepted for publication in the Journal of Clinical Epidemiology.

Chapter 3.2
K.J.M. Janssen, C.J. Kalkman, G.J. Bonsel, D.E. Grobbee, K.G.M. Moons, Y. Vergouwe.
Validation of a clinical prediction rule for severe postoperative pain in new settings.
In s*econd revision for Anesthesia&Analgesia*.

Chapter 3.3
K.J.M. Janssen, Y. Vergouwe, C.J. Kalkman, D.E. Grobbee, K.G.M. Moons.
A simple method to update clinical prediction rules.
Submitted.

Chapter 3.4
K.J.M. Janssen, K.G.M. Moons, C.J. Kalkman, D.E. Grobbee, Y. Vergouwe.
Updating a clinical prediction model improved the performance in new patients.
Journal of Clinical Epidemiology, in press.

Chapter 4.1
K.J.M. Janssen, Y. Vergouwe A.R.T. Donders, F.E. Jr. Harrell, Q. Chen, D.E. Grobbee, K.G.M. Moons.
Clinical prediction models in daily practice: Practical solutions for missing predictor values.
Submitted.

Contents

Chapter 1

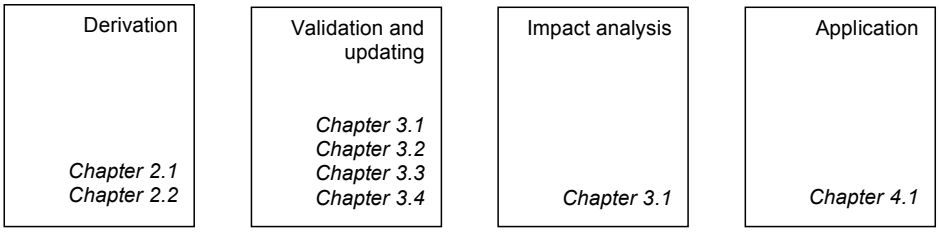
Introduction

In clinical prediction research, patient characteristics, test results and disease characteristics are often combined to estimate the risk that a disease or outcome is present (diagnosis) or will occur (prognosis) in so-called prediction models. In daily practice physicians often make implicit risk estimations using their clinical knowledge and experience. Prediction models are meant to make these estimations more explicit. The number of clinical prediction models has sharply increased in the literature; from 6,744 published articles in 1995 to 15,450 in 2005 using a suggested search-strategy in Medline[1].

The predicted risks of the models can be used for several purposes, for example, to inform patients and their families, or to assist in clinical decisions for individual patients, such as treatment choices. Also, prediction models can be used to stratify patients by disease severity and create risk groups for specific selection in randomized therapeutic trials. Moreover, the models can be used as quality of care instruments, of e.g. hospitals. In that case, the predicted risks can be compared with the observed actual outcomes in a particular hospital. For example, when the observed mortality in neonatal intensive care units is lower than the predicted mortality as estimated by the CRIB (clinical risk index for babies) score[2], the neonatal care of the unit is above average.

Roughly four phases can be distinguished in clinical prediction research; derivation of the prediction model, validation of the model in new subjects and updating (if necessary), assessment of its clinical impact, and the application of the clinical prediction model in daily practice (Figure 1).

Figure 1 Consecutive phases in prediction research



Given the commonly applied dichotomization of diagnostic and prognostic outcomes - presence or absence of a particular disease or outcome - many prediction models are derived with dichotomous logistic regression analysis.[3;4] Other methods to derive prediction models for dichotomous outcomes are for example classification and regression trees (CART) and artificial neural networks. However, it has repeatedly been shown that both methods no not produce prediction models that achieve higher predictive performance than models derived by multivariable logistic regression.[5-9] Genetic programming is a more novel and promising technique that may require further investigation.[10-13]

The predictive performance of most derived prediction models is decreased when tested in new patients (Figure 1**,** phase 2), compared to the performance estimated in the patients that were used to derive the model (Figure 1**,** phase 1).[3;14;15] Therefore, before a prediction model can be applied in daily clinical practice (Figure 1**,** phase 4), it is widely suggested that any prediction model needs to be tested (i.e. externally validated) in new patients (Figure 1**,** phase 2). However, when the predictive performance is disappointing

in the validation data set, the original prediction model is frequently rejected and the researchers simply pursue to build their own (new) prediction model on the data of their patients. When every new patient sample would lead to a new prediction model, the prior information that is captured in previous studies and prediction models would be neglected. Moreover, it would lead to many prediction models for the same outcome, obviously creating impractical situations as physicians have to decide on which model to use. For example, there are over 60 published models to predict outcome after breast cancer [16]. The alternative to redevelopment of a new model on every other data set, is updating existing prediction models. The updated models combine the information that is captured in the original model with the information of the new patients.[17-20] As a result, updated models are adjusted to the new patients and thus based on data of the original and new patients, potentially increasing their generalisability.

This thesis aims to improve methods of clinical prediction research, with a focus on the derivation and validation plus updating of prediction models (see Figure 1). The first part addresses the derivation phase in prediction research (Figure 1, phase 1). Many guidelines regarding the derivation phase have indeed been published[3;4;21]. However, dealing with missing values is an underappreciated aspect in this derivation phase, let alone in the other three phases of prediction research. Researchers often simply neglect the data of patients with missing data records, because this is what standard software packages do when the data is analyzed (complete case analysis). As this leads to a smaller dataset, it comes at least at the price of loss of power. As an alternative, researchers tend to drop a variable from the analysis when it has missing data. However, both methods may do more harm than good by leading to biased results and potentially to suboptimal patient care.[22-31] Multiple imputation is more sophisticated method to deal with missing data. It is a statistical technique that uses all other observed variables to fill in logical values for the missing data.[24;26;29-36] Chapter 2.1 compares three methods that can handle missing predictor values when a prediction model is derived: complete case analysis, dropping the predictor with missing values and multiple imputation. The percentage of missing values in one of the predictors ranges between 10% and 90%. In Chapter 2.2, we compare alternative methods to derive prediction models. We compare the predictive performance of four different approaches. Three models were derived with logistic regression (by logistic regression without shrinkage, logistic regression with a single shrinkage factor, or by logistic regression with inherent shrinkage by penalised maximum likelihood estimation), and the fourth model was derived by genetic programming.

The second part of this thesis concerns the validation phase of prediction research (Figure 1**,** phase 2). Chapter 3.1 presents an overview of the different types of validation studies that can be distinguished, namely temporal, geographical, and domain validation. Solutions to improve or update a model in case of disappointing performance in a validation study are discussed, and impact studies (phase 3, Figure 1) are described as well. In Chapter 3.2, we validated a prediction model that preoperatively predicts the risk of severe pain in the first postoperative hour in surgical patients. The rule was validated in patients that underwent surgery later in time in another hospital. In Chapter 3.3, we show that a simple updating strategy, that combines the information of the original model with the information of the new patients, can substantially improve the predictive performance of a prediction model in new patients and is preferred over deriving a new model. In Chapter 3.4, we discuss five updating methods that vary in extensiveness, which is

reflected by the number of model parameters that is adjusted or re-estimated. We present an empirical example and show the results of simple and more extensive updating methods.

The third part of this thesis addresses the application of a prediction model (Figure 1**,** phase 4). In Chapter 4.1, we study the situation in which a physician applies a prediction model for a patient in which certain test results or other predictor values were not available (missing). We present six strategies that handle such missing predictor values in clinical practice and compare the effects on the predictive performance of the prediction model.

This thesis ends with an overview of the promises and pitfalls of using the electronic patient records (EPR) as a basis for prediction research to enhance patient care, and vice versa. The EPR are medical records in digital format that facilitate storages and retrieval of data on patient care. Though the primary aim of the EPR is to aid patient care it creates highly attractive opportunities for research, notably with regard to diagnostic and prognostic prediction studies.

References

1 Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001; 8(4):391-397.

2 The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. The International Neonatal Network. Lancet 1993; 342(8865):193-198.

3 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(4):361-387.

4 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997; 277(6):488-494.

5 Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Investig Med 1995; 43(5):468-476.

6 Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. 2001. J Clin Epidemiology. Ref Type: Generic

7 Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. 1996. J Clin Epidemiol. Ref Type: Generic

8 Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Medinfo 1998; 9 Pt 1:493-7.:493-497.

9 Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998; 17(21):2501-2508.

10 Biesheuvel CJ, Siccama I, Grobbee DE, Moons KG. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. J Clin Epidemiol 2004; 57(6):551-560.

11 Forrest S. Genetic algorithms: principles of natural selection applied to computation. Science 1993; 261(5123):872-878.

12 Podbregar M, Kovacic M, Podbregar-Mars A, Brezocnik M. Predicting defibrillation success by 'genetic' programming in patients with out-of-hospital cardiac arrest. Resuscitation 2003; 57(2):153-159.

13 Tsakonas A, Dounias G, Jantzen J, Axer H, Bjerregaard B, von Keyserlingk DG. Evolving rule-based systems in two medical domains using genetic programming. Artif Intell Med 2004; 32(3):195-216.
14 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130(6):515-524.
15 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006; 144(3):201-209.
16 Altman DG. Breast cancer. Translational therapeutic strategies. New York: 2007.
17 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23(16):2567-2586.
18 Hosmer D.W., Lemeshow S. Applied logistic regression. New York: John Wiley and Sons, Inc.; 1989.
19 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999; 99(16):2098-2104.
20 van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995; 14(18):1999-2008.
21 Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54(8):774-781.
22 Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003; 56(1):28-37.
23 Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995; 48(2):209-219.
24 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12):1255-1264.
25 Little R.J. Regression with missing X's: a review. J Am Stat Assoc 1992; 87:1227-1237.
26 Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, New Jersey: John Wiley & Sons; 1987.
27 Little RJ. Methods for handling missing values in clinical trials. J Rheumatol 1999; 26(8):1654-1656.
28 Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, New Jersey: John Wiley & Sons; 1987.
29 Rubin DB. Multiple Imputation after 18+ years. Journal of the American Stat Association 1996; 91:473-489.
30 Schafer JL. Analysis of Incomplete Multivariate Data. Chapman & Hall /CRC; 1997.
31 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7(2):147-177.
32 Barnes SA, Lindborg SR, Seaman JW, Jr. Multiple imputation techniques in small sample clinical trials. Stat Med 2006; 25(2):233-245.
33 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10):1087-1091.
34 Harrell FE, Jr. Regression modelling strategies. Springer-Verlag, New York; 2001.
35 Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10):1092-1101.
36 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10(4):585-598

Chapter 2

Derivation

Chapter 2.1

Missing data in medical research: making up is better than giving up

Abstract

Background No matter how hard researchers try to prevent it, missing data always occur in medical research. Researchers commonly neglect missing data (complete case analysis), or drop variables with missing values from the analysis. It can lead to loss of power and biased results. Multiple imputation of missing values is a more sophisticated method. We compared these three methods for dealing with missing values.

Method We used the data from a cross sectional study that aimed to predict the presence of deep venous thrombosis (DVT). 804 consecutive patients presented themselves to the physician with a suspicion of DVT. We selected three independent variables: D-dimer level, calf circumference and history of leg-trauma. We assigned missing values, ranging from 10% to 90%, to one independent variable and applied the three methods. The data were further analysed using multivariable logistic regression, which was repeated 500 times using simulations. The outcome of interest was the bias in the effect estimate (regression coefficient) of independent variables, coverage of corresponding 90% confidence interval, and ability of the model to discriminate between presence and absence of DVT (ROC area).

Results Multiple imputation outperformed complete case analysis and dropping the variable with missing data from the analysis for bias and coverage. The ROC area after multiple imputation was unbiased, but substantially decreased after complete case analysis or dropping the variable with missing data.

Interpretation Multiple imputation is to be preferred over dropping the variable with missing data or complete case analysis, irrespective of the percentage of missing values.

Introduction

No matter how hard researchers try to prevent it, missing data always occur in any kind of medical research. Commonly, researchers simply neglect the data of patients with missing data records, because this is what standard software packages do when the data is analyzed (complete case analysis). As this leads to a smaller dataset, it comes at least at the price of loss of power. As an alternative, researchers tend to drop a variable from the analysis when it has missing data. However, an even higher price has to be paid when the variable with missing data is dropped from the analysis or when a complete case analysis is performed, as both methods may do more harm than good by leading to biased results and potentially to suboptimal patient care.[1-10]

How this bias is reflected in the results depends on the type of medical research. In diagnostic or prognostic prediction studies the focus can be on the predictive ability of a single diagnostic or prognostic factor. However, in many prediction studies patient characteristics (demographic factors, patient history, physical examination, and additional test results) are combined to develop a so-called multivariable prediction model or risk score, to study the simultaneous predictive effect of several predictors on the outcome. Such a prediction model aims to predict the risk of a diagnostic or prognostic outcome for individual patients. For example, one may study the predictive effect of body mass index (BMI), age, gender, the intake of saturated fat and other life style factors on the risk of cardiovascular diseases (CVD) for individual people. Interest is in the regression coefficients of the predictors, and on the discriminative ability of the prediction model, i.e. the ability of the model to distinguish between patients at high and low risk of CVD.

A complete case analysis or dropping the predictor with missing data from the analysis and thus from the prediction model may result in biased regression coefficients of the other predictors and on a discriminative value that is too low, since the predictive effect of one of the predictors is not taken into account.

In etiologic (or causal) studies, one rather studies the effect of a specific etiologic factor on the outcome of interest, corrected for the influence of confounders. Following the previous example, the effect of BMI on the risk of CVD could be studied, corrected for the influence of age, gender, the intake of saturated fat and other life style factors. When there is missing data in one of the confounders, for example the intake of saturated fat, complete case analysis or dropping this confounder from the analysis may lead to an erroneous estimation of the effect of BMI on the risk of CVD.

In contrast to the ad hoc methods complete case analysis and dropping the variable from the analysis, multiple imputation is more sophisticated method to deal with missing data. It is a statistical technique that uses all other observed variables to fill in plausible values for the missing data, and that receives increasing attention in the medical literature.[1;3;5;7-19] However, many researchers are unaware or uncertain about this approach to handle missing data and still drop variables with missing data from the analysis or perform a complete case analysis. Complete case analysis not always necessarily leads to biased results. Under the condition that the missing data are Missing Completely at Random (MCAR), meaning that the cause of missingness is pure coincidence, complete case analysis will not lead to biased results. However, it is widely acknowledged that missing data in medical studies are seldom MCAR but mostly related to other patient characteristics, called Missing At Random (MAR)[9]

We propose it is preferable to impute variables with missing values by multiple imputation rather than to drop them, even when the percentage of missing data is high. We compare the effects of multiple imputation, complete case analysis and dropping the variable from the analysis on the bias and discriminative ability of a multivariable model when the amount of missing data in a variable ranges from 10 percent to 90 percent. We use empirical data of a study in which a prediction model was developed for the detection of deep venous thrombosis (DVT). We also discuss the impact of our methods for etiologic studies. Although this data originates from a cross sectional study, the results are generalisable to follow-up data as well.

Methods

Empirical data

Data were obtained from a large prospective, cross sectional study among adult patients with a suspicion of deep venous thrombosis (DVT). For specific details and main results of the study we refer to the literature.[20-22] In brief, consecutive patients with a suspicion of DVT who visited one of the 110 participating primary care physicians adherent to three non-academic hospitals in the Netherlands were included. Suspicion of DVT was primarily based on the presence of at least one of the following symptoms or signs of the lower extremities that existed less than 30 days: swelling, redness, or pain in one of the legs. After informed consent, the primary care physician systematically documented the patient's history and physical examination. Subsequently, venous blood was drawn to measure the D-dimer level using the quantitative ELISA method and the latex assay method. Finally, all patients were referred to the hospital to undergo the reference test – i.e. repeated compression ultrasonography of the lower extremities – to determine the presence or absence of DVT. The observer of the reference test was blinded to the results of the patient history, physical examination and the D-dimer level.

For our illustration, we specifically selected two dichotomous variables (history of a leg trauma and a difference in calf circumference of 3 cm or more) and one continuous variable (D-dimer level) with different correlations. A difference in calf circumference of 3 cm or more was associated with the D-dimer level (correlation coefficient $r = 0.28$), while history of a leg trauma was not associated with the D-dimer level ($r = 0.04$) nor with a difference in calf circumference of 3 cm or more ($r = 0.06$). All three variables were predictors of the presence of DVT.

For this illustration 804 patients that had no missing data on any of these three variables were included, of whom 160 had DVT (prevalence = 20 percent). This will be called the original study sample from now on. Seventeen percent had a leg trauma in the past 4 weeks and 38 percent had a difference in calf circumference of 3 cm or more (Table 1).

We fitted a multivariable logistic regression model with the three variables as the independent variables and DVT presence/absence as the outcome. The natural logarithm of the D-dimer level was used as this increased the predictive ability of the model. The estimated regression coefficients in the original study sample were considered as the 'true' values, to which all subsequent estimations were compared.

Table 1 Distribution of the studied predictors: the (natural logarithm of) the D-dimer level, history of a leg trauma (yes/no) and difference in calf circumference of 3 cm or more (yes/no), and the true values of the logistic regression coefficients.

Predictors	Distribution % (n)	True regression coefficients[a]
Intercept	-	-13.24
History of a leg trauma	17 (136)	-0.50
Natural logarithm of the D-dimer level[b]	6.83 (1.49)	1.58
Difference in calf circumference of 3 cm or more	38 (306)	0.60

[a] No 95% confidence interval is given as these regression coefficients are considered to be the truth
[b] Mean (standard deviation)

Missing data

Missing data were introduced in one variable, the D-dimer level. Missing data can be caused by several mechanisms, for example by random factors, when a tray with blood samples drops from a table and the samples can therefore not be analyzed. This is called missing completely at random (MCAR).[7] However, often the missingness is related to other observed patient characteristics. For example, patients that are relatively healthier might be less likely to undergo subsequent, more invasive tests, leading to more missing data on those tests for these patients. Such missing data are called Missing At Random (MAR).[7] The missing data are random *conditional* on the other available information. If there is no information about the reason for missingness, these missing data are called Missing Not At Random (MNAR) or non-ignorable missing. This means that the probability that an observation is missing depends on unobserved subject information. In medical research, missing data are commonly not MCAR or MNAR, but related to other observed subject characteristics, i.e. MAR.[9;10] We therefore generated the missing data in the D-dimer level according to a MAR mechanism. Hence, the probability of missing data depended on the other observed variables, namely the presence of DVT and a difference in calf circumference of 3 cm or more. The missing data in the D-dimer level were generated with percentages ranging between 10 percent and 90 percent, in steps of 10 percent.

More information about the mechanism that was used to create missing data is provided in the appendix. All simulations were performed using R2.3.1 (R Foundation for Statistical Computing, www.R-project.org). The simulation script is available on request. We used the original study sample (804 patients) as the basis for 500 simulated datasets, in which we repeated the mechanism of introducing missing data. Because regression analysis is conditional on the values of the independent variables (and not on the outcome or dependant variable), we used the values of the three variables to simulate all 500 datasets. Therefore, in each of the 500 datasets the variable values for all 804 patients remained identical to the values in the original study sample. Only the presence or absence of DVT was simulated. This was done by using the true regression coefficients (Table 1) of the variables and the patient's values to calculate the true probability of DVT (P) for each patient. Next, for each patient a random number from a uniform distribution in the interval [0,1] was sampled. An event status of 1 (DVT present) was assigned when the random number was smaller than P and an event status of 0 otherwise. For example, for patients with a DVT probability of 0.3, 30% would have DVT and been assigned an outcome value of 1. Thus, the outcome values but not the variable values differed in the 500 simulated datasets.

Methods to handle missing data

In all 500 datasets, we used three methods to deal with the missing data before we fitted the multivariable logistic regression model. Each method was applied to all simulations resulting in 500 fitted models per method. Per method, the 500 estimated regression coefficients and standard errors were summarized. The three methods were:

1. Complete case analysis: analyzing only the subjects without missing data.
2. Dropping the variable with missing data (D-dimer level): include only the remaining variables without missing data in the analysis (a difference in calf circumference of 3 cm or more and a history of a leg trauma).
3. Multiple imputation using *mice* (S. Van Buuren & C.G.M. Oudshoorn, http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm). Multiple imputation replaces reach missing value with 10 values drawn from an appropriate estimated distribution. Per simulation, the imputation model was estimated and 10 imputed datasets were created, which were all analyzed using the same standard method, i.e. fitting the previously mentioned multivariable logistic model. The models were combined using standard statistical methods in a way that reflects the extra variability due to missing data.[7]

Outcome measures

We compared the results of the three methods to the results of the analyses on the original study sample. The outcome measures of interest were the bias of the regression coefficients of the variables, coverage of the confidence intervals of the regression coefficients and the discriminative ability of the model.

1. Bias. We calculated the discrepancy between the true value of the regression coefficients of the three variables (Table 1) and the corresponding mean of the 500 estimations of these regression coefficients in the simulated data.
2. Coverage of confidence intervals. We estimated the percentage (over the 500 simulations) of the 90 percent confidence intervals (90 percent CI) that indeed included the true value of the regression coefficient of each variable. Values near 90 percent represent adequate coverage, values < 90 percent indicate that the 90 percent CI is too narrow. This implies that in studies where the null hypothesis is valid ('variable has no effect') too often a significant effect will be found. Values > 90 percent indicate that the 90 percent CI is too wide, which implies that the power of the study is suboptimal and more type II errors may occur.
3. Discriminative ability, expressed by the Receiver Operating Characteristic (ROC) curve area or the c-statistic.[23;24] The ROC area is a commonly used measure to indicate the overall discrimination, i.e. the ability of a prediction model to distinguish between patients with and patients without DVT. An ROC area ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination). For each fitted model, the ROC area was estimated and averaged over the 500 simulations. We calculated the discrepancy between the true value of the ROC area (0.88, 95% confidence interval 0.82-0.94) and the average ROC area of the fitted models in the 500 simulations.

Etiologic perspective

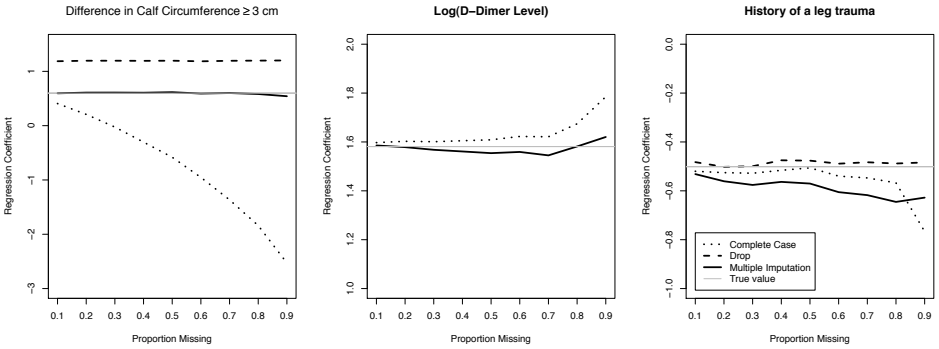
Obviously, none of the three variables in our data were (potential) etiologic factors of DVT. However, statistically we could analyse our data as if it was an etiologic study. As a result, we can illustrate the effects of a complete case analysis, dropping variables (confounders) with missing values from the analysis and multiple imputation, in an etiologic context with missing values on a confounder variable, a frequently encountered situation. To do so, we hypothetically defined a difference in calf circumference of 3 cm or more as the etiologic factor, with a history of a leg trauma and the D-dimer level as confounders. The confounder D-dimer level had missing values ranging from 10% to 90% as described above. We used the same data (rather than introducing another data set) only to enable generalisation of our results both to a prediction and an etiologic context.

Results

Bias

Figure 1 shows the discrepancies (i.e. bias) between the mean of the 500 simulation estimates and the true value of the regression coefficients of the three variables Complete case analysis resulted in a severely biased (underestimated) regression coefficient of a difference in calf circumference of 3 cm or more. The bias sharply increased when the percentage of missing data in the D-dimer level increased. Dropping the D-dimer level from the analysis led to a biased (systematically overestimated) regression coefficient of difference in calf circumference, irrespective of the percentage of missing data in the D-dimer level. Note that any variability in the estimates of the regression coefficients when the predictor with missing values is dropped across different percentages of missing values is due to sampling variation in the 500 simulations. Multiple imputation led to an unbiased regression coefficient of difference in calf circumference. For the D-dimer level itself, complete case analysis resulted in a biased regression coefficient only when the percentage of missing data was high (over 70%).

Figure 1 Estimates of the regression coefficients of the difference in calf circumference of 3 cm or more, the natural logarithm of D-dimer level and a history of a leg trauma, after applying complete case analysis (dotted line), dropping the D-dimer level from the analysis (dashed line) and multiple imputation (black solid line). The percentage of missing data in D-dimer level increased from 10 percent to 90 percent, in steps of 10 percent. The grey solid line represents the true value of the regression coefficients.



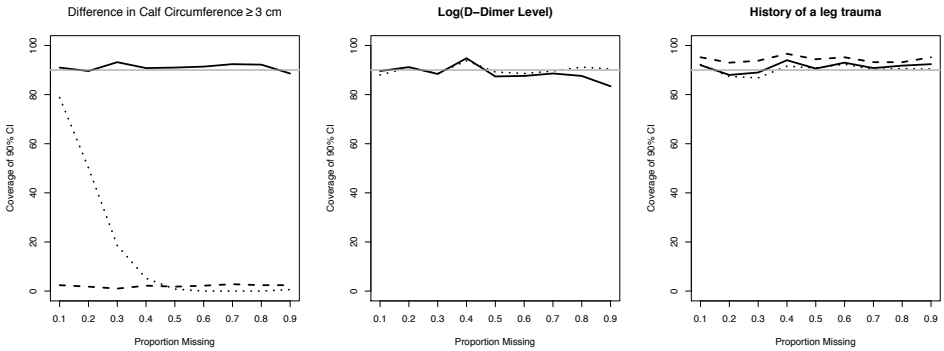
However, this bias was small, as the true value was -1.58 (Table 1) and the most severely biased estimate -1.78. Multiple imputation led to an unbiased regression coefficient of the D-dimer level. Logically, the regression coefficient of the D-dimer level could not be estimated when it was dropped from the analyses.

Complete case analysis resulted in a biased regression coefficient of a history of a leg trauma only when the percentage of missing data was about 60% or higher. Dropping the D-dimer level did not lead to a biased estimate of a history of a leg trauma. Multiple imputation led to a biased estimate of a history of a leg trauma when the percentage of missing values increased. However, this bias was small, as the true value was -0.50 and the most severely biased estimate -0.65.

Coverage of the 90 percent confidence intervals

Figure 2 shows the coverage of the 90 percent CI of the regression coefficients. Complete case analysis led to a coverage of a difference in calf circumference that was too low, and that substantially decreased when the percentage of missing data in the D-dimer level increased. Dropping the D-dimer level from the analysis led to a coverage of almost zero, irrespective of the percentage of missing data. Multiple imputation led to good coverage (around 90%) of the 90 percent confidence interval of the regression coefficient of a difference in calf circumference. Complete case analysis and multiple imputation led to good coverage of the 90 percent confidence interval of the regression coefficient of the D-dimer level. All three methods led to good coverage of the 90 percent confidence interval of the regression coefficient of a history of a leg trauma.

Figure 2 Coverage of the 90 percent confidence intervals (CI) of the regression coefficients of a difference in calf circumference of 3 cm or more, the natural logarithm of D-dimer level and a history of a leg trauma, after applying complete case analysis (dotted line), dropping the D-dimer level from the analysis (dashed line) and multiple imputation (black solid line). The percentage of missing data in the D-dimer level increased from 10 percent to 90 percent, in steps of 10 percent. The grey solid line represents the true value of the coverage.



Discriminative ability

Figure 3 shows the ROC area of the multivariable (prediction) model. The ROC area after complete case analysis dropped from 0.88 (true value) to 0.77 when the percentage of missing data increased from 30% to 90%. The ROC area was severely decreased (0.65) when the D-dimer level was dropped from the analysis. The ROC area of the model after multiple imputation was equal to the true discriminative ability.

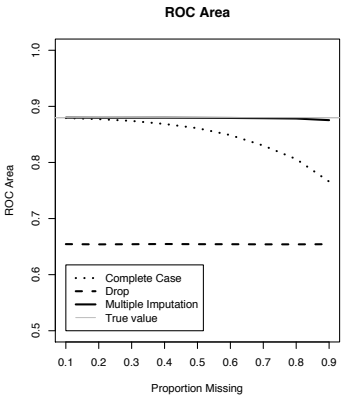


Figure 3 Receiver Operating Characteristic (ROC) curve area of the prediction model, after applying complete case analysis (dotted line), dropping the predictor with missing data from the analysis (dashed line) and multiple imputation (black solid line), when the percentage of missing data in the predictor increased from 10 percent to 90 percent, in steps of 10 percent. The grey solid line represents the ROC area when there were no missing data (true value).

Etiologic perspective

When we would (hypothetically) interpret the same data as if it was an etiologic study, with the missing values occurring in a confounder (here: D-dimer level), we would draw the same conclusions. If difference in calf circumference would be the etiologic factor, complete case analysis and simply dropping the confounder from the analyses resulted in a biased regression coefficient of the etiologic factor with substantially decreased coverage across all missing confounder value percentages. In contrast, multiple imputation would yield results, both in terms of bias and coverage, that were close to the truth.

Discussion

In a simple multivariable diagnostic study (with three variables), we compared multiple imputation with complete case analysis and dropping a variable from the analysis when the amount of missing data in that variable ranged between 10 percent and 90 percent.

Complete case analysis led to two slightly biased regression coefficients (only when the percentage of missing data in the D-dimer level was high) and one severely biased regression coefficient. Dropping the D-dimer level from the analysis led to one severely biased regression coefficient and one unbiased regression coefficient of the two remaining variables. Multiple imputation led to two unbiased and one slightly biased regression coefficient. Complete case analysis led to good coverage of the confidence interval of two regression coefficients but for one it was too low. Dropping the D-dimer level from the analysis led to good coverage of the confidence interval of one regression coefficient but to a coverage close to zero for another. Multiple imputation led to good coverage of the 90 percent confidence interval of all three regression coefficients, irrespective of the percentage of missing data.

Also when the data are hypothetically interpreted from an etiologic point of view, with missing data occurring in a confounder (a common situation), similar inferences would be made. Multiple imputation leads to less biased regression coefficients of the etiologic factor and proper coverage of the corresponding confidence interval, also when the percentage of missing values in the confounder is high, compared to complete case analysis or dropping the confounder from the analysis.

In multivariable diagnostic (or prognostic) prediction research, the main outcome measure commonly is the discriminative ability of the developed prediction model. Complete case analysis led to a decreased discriminative ability when the percentage of missing data increased. The discriminative ability was severely decreased when the predictor with missing data was dropped from the analysis, irrespective of the percentage of missing data. However, the discriminative ability after multiple imputation remained unchanged.

Coverages of the 90 percent CI that are too low can lead to erroneous inferences. For example, when the null hypothesis ('etiologic factor has no effect: regression coefficient = 0') is true and the coverage of the confidence interval is 60 percent instead of 90 percent, the null hypothesis will be wrongly rejected in 40 percent (instead of 10 percent) of the time. This means that in 40 percent (instead of 10 percent) of the time, one concludes that the etiologic factor has an effect on the outcome, while in reality there is none.

Following many others, we found that a complete case analysis provides severely biased results.[1-9] Even for high percentages of missing data, multiple imputation provided more valid (less biased) results than dropping the variable from the analysis. Yet, we should not focus on the extreme high percentages of missing data, as we by no means want to suggest that it is legitimate to analyses datasets with up to 90% missing values. The percentage of missing values ranging from 10% to 90% only illustrates that the impact of multiple imputation is independent of the percentage of missing values. Also, our original study sample consisted of 804 patients. When 90 percent of the values of a predictor was missing, data of 80 patients was left that could be used in the analysis. However, we do not know how the results would be if the original study sample consisted of less patients. For instance, in a study sample of 200 patients in which 90 percent of the values of one variable is missing, the data of only 20 patients can be used for the analyses. Therefore, one should not interpret the results of our study as a statement that multiple imputation can handle up to 90 percent missing data in all instances. One should realize that it is not so much the percentage of missingness that is of importance, but rather the amount of data that is left for the analyses.

In our hypothetical exercise to interpret the same data from an etiologic perspective, we first considered difference in calf circumference as the etiologic factor, with the D-dimer level (in which missing values were assigned) and history of a leg trauma as confounders. The regression coefficient of history of a leg trauma was not biased when the data of the patients with missing data (D-dimer level) was dropped from the analysis. Hence, had we chosen history of a leg trauma as the etiologic factor (and a difference in calf circumference and the D-dimer level as the confounders), the inferences seem to be different and not per se in favour of multiple imputation. However, the bias in the regression coefficient of history of a leg trauma after multiple imputation was relatively small: a deviation of -0.50 (true value) to -0.65 (most severely biased estimate). Furthermore, strictly speaking the D-dimer level would not be a confounder of the 'causal' association between a history of a leg trauma and the presence of DVT. Namely, the D-dimer level is hardly associated with a history of a leg trauma (correlation = 0.04), while an association with the etiologic factor (history of a leg trauma) is a prerequisite for a confounder (D-dimer level). Finally, when we repeated the analyses simulating the missing data in a difference in calf circumference of 3 cm or more instead of in the D-dimer level, multiple imputation again showed better results than complete case analysis or dropping the variable with missing data from the analysis (data not shown).

Dropping the confounder with missing data will not always result in biased estimates of the other confounders. In our study, the regression coefficient of the other confounder history of a leg trauma was not biased. One could suggest that there might be situations in which neither the regression coefficient of the etiologic factor will be biased when the confounder with missing data is dropped from the analysis. However, history of a leg trauma was not associated with the D-dimer level, and therefore dropping the D-dimer level from the analysis did not lead to a different association between history of a leg trauma and the presence of DVT. Therefore, although dropping the confounder with missing data did not always lead to a biased regression coefficient of the other confounder, this does not suggests that in some situations it is safe to drop confounders from the analysis. Namely, one does not always know beforehand whether a confounder is (strongly) related to other confounders. It is therefore better to use a safe strategy and multiple impute the missing data than to do a complete case analysis or to drop the confounder from the analysis.

We simulated the missing values in our dataset according to a MAR mechanism. One may suggest that this starting point may not be correct in every situation. However, it is known from the literature that missing data in medical or social sciences often occur on a MAR mechanism[7;9;10]. One may criticize this assumption, and address the question whether one can be sure that the missing data are MAR and not MNAR. This is a serious issue that should not be set aside too easily. However, it has been shown that even when missing data are not precisely MAR, multiple imputation still tends to do better than ad hoc methods like dropping the variable with missing data or performing a complete case analysis.[9] Besides that, it has also been shown that in many realistic cases an erroneous assumption of MAR will often have only a minor impact on the results, especially in datasets that contain many detailed patient characteristics.[10;25] Yet, there are also situations in which an erroneous assumption of MAR will have an impact on the results.[10]

We used multiple imputation to handle the missing data in our study. Another advocated method to handle missing data is the maximum likelihood estimation (e.g. using the EM-algorithm). However, maximum likelihood estimations are particularly applied in multilevel or repeated-measurement analysis in which variables are documented more than once. In this study, we focus on a situation where the variables and the outcome are measured once, for which multiple imputation is the advocated method.[4;5;7-10;15;18;19] We used the software package Mice for multiple imputation in this study. Several other packages exist. For a comparison of frequently used software packages for multiple imputation, we refer to the literature.[26;27]

In agreement with previous studies, we found that complete case analysis leads to biased study results. Besides that, our results show that it is not harmless to simply drop a covariate or confounder with (many) missing data from the analysis, as this may lead to seriously biased estimates of the effect of the etiologic factor under study. Further, the discriminative ability of a developed multivariable prediction model can be severely reduced when a predictor with missing data is dropped from the analysis. Ironically, the greater the extent of missing data the better the relative performance of multiple imputation. That is not because it is good to have more missing data but rather because simplified alternatives to dealing with missing data are deficient. Although we by no means want to suggest that less effort should be undertaken to collect as many data as possible in a research setting, in reality every researcher faces the problem of missing data, irrespective of the

efforts that are undertaken to avoid it. For those situations, it is better to 'make up' data according to the strict methodology of multiple imputation than the alternative, which is to give up or delete real data. Therefore, we conclude that multiple imputation should be preferred over dropping the variable with missing data or conducting a complete case analysis, when handling missing data in medical research.

References

1 Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003; 56(1):28-37.
2 Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995; 48(2):209-219.
3 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12):1255-1264.
4 Little R.J. Regression with missing X's: a review. J Am Stat Assoc 1992; 87:1227-1237.
5 Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, New Jersey: John Wiley & Sons; 1987.
6 Little RJ. Methods for handling missing values in clinical trials. J Rheumatol 1999; 26(8):1654-1656.
7 Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, New Jersey: John Wiley & Sons; 1987.
8 Rubin DB. Multiple Imputation after 18+ years. Journal of the American Stat Association 1996; 91:473-489.
9 Schafer JL. Analysis of Incomplete Multivariate Data. Chapman & Hall /CRC; 1997.
10 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7(2):147-177.
11 Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res 1999; 8(1):17-36.
12 Barnes SA, Lindborg SR, Seaman JW, Jr. Multiple imputation techniques in small sample clinical trials. Stat Med 2006; 25(2):233-245.
13 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10):1087-1091.
14 Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidemiol 2002; 55(2):184-191.
15 Harrell FE, Jr. Regression modelling strategies. Springer-Verlag, New York; 2001.
16 Kmetic A, Joseph L, Berger C, Tenenhouse A. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. Epidemiology 2002; 13(4):437-444.
17 Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10):1092-1101.
18 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10(4):585-598.
19 Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999; 8(1):3-15.
20 Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005; 94(1):200-205.
21 Oudega R, Moons KG, Hoes AW. Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care. Fam Pract 2005; 22(1):86-91.
22 Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous

thrombosis in primary care patients. Ann Intern Med 2005; 143(2):100-107.
23 Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148(3):839-843.
24 Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3(2):143-152.
25 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6(4):330-351.
26 Horton NJ, Lipsitz SR. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. The American Statistician 2001; 55(3):244-254.
27 Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 2007; 61(1):79-90.

Appendix

Missing data were generated using a missing at random (MAR) strategy: the probability that an observation was missing was dependent on the other observed patient characteristics. The introduction of missing data in the D-dimer level depended on the presence of DVT and a difference in calf circumference of 3 cm or more. Therefore, we could distinguish four groups; patients without DVT with a difference in circumference of 3 cm or more, patients without DVT with a difference in circumference of 3 cm or more, patients with DVT with a difference in circumference of 3 cm or more, and patients with DVT with a difference in circumference of 3 cm or more. The number of observations in each group depended on the simulated dataset, since the presence of DVT was newly generated in each simulation. The odds ratios of a missing response are presented in Table A. The advantage of using odds ratios is that these can be scaled up and down based on the number of missing data that need to be generated. All simulations were performed using R2.3.1 (R Foundation for Statistical Computing, www.R-project.org). The simulation scripts are available on request.

Table A odds ratios of a missing value in the D-dimer test result

	DVT not present	DVT present
Difference in calf circumference < 3 cm	16	1
Difference in calf circumference ≥ 3 cm	5	10

Chapter 2.2

Derivation methods for clinical prediction models: a comparison between conventional logistic regression, penalised maximum likelihood estimation and genetic programming.

Abstract

Background During the derivation of prediction models, the predictive strength of candidate predictors is assessed, and the need for transformations of continuous predictors is tested in the dataset at hand. A disadvantage of using the data for the selection of predictors and testing for transformations is that the models can be overfitted. During or after derivation of the models, the regression coefficients can be corrected for overfitting by shrinkage. We compared the accuracy of a model that predicts the presence or absence of Deep Venous Thrombosis (DVT) when derived by four different methods.

Methods We used the data of a cohort of 2086 primary care patients suspected of DVT (415 patients with DVT), which included 21 candidate predictors. The cohort was split into a derivation set (1668 patients, 329 with DVT) and a validation set (418 patients, 86 with DVT). Also, 100 cross validations were conducted in the derivation set. The models were derived by logistic regression, logistic regression with shrinkage by bootstrapping techniques, logistic regression with shrinkage by PMLE (a method in which the individual regression coefficients are separately adjusted for overoptimism), and genetic programming (a search method inspired by the biological model of evolution). The accuracy of the models was tested by assessing the discrimination and calibration.

Results There were only marginal differences in the discrimination and calibration of the models in the 100 cross validations, and in the validation set.

Discussion The accuracy measures of the models derived by logistic regression, logistic regression with shrinkage by bootstrapping techniques, logistic regression with shrinkage by PMLE, and genetic programming were only slightly different, and the 95% confidence intervals mostly overlapped. The choice between these derivation methods should be based on the characteristics of the data and situation at hand.

Introduction

In clinical prediction research, patient characteristics, test results and disease characteristics are often combined in so-called prediction models to estimate the risk that a disease or outcome is present (diagnosis) or will occur (prognosis).[1;2] During the derivation of these prediction models, the predictive strength of candidate predictors is assessed, and the need for transformations of continuous predictors is tested. Often only the strongest predictors are included in the final model, since parsimonious models are easier to apply in daily clinical practice. A disadvantage of using the data for the selection of predictors and testing for transformations is that the models can be overfitted (especially in relatively small datasets). [3-6] The overfitted regression coefficients result in too extreme predictions (i.e. the low probabilities predicted too low and the high probabilities predicted too high). Moreover, the estimated predictive accuracy of the model is likely to be too optimistic, implying that the accuracy in new patients will be decreased.

The risk of overfitting can be reduced by restricting the selection of predictors on the data. A rule of thumb is that at least 10 events are needed to consider one degree of freedom.[5;7] For example, to derive a prediction model in a dataset of 500 patients of which 200 experience the outcome of interest, maximal 20 predictors (with 1 degree of freedom) can be considered for inclusion. When the transformation of continuous predictors is tested, these degrees of freedom should also be taken into account in this maximum number of degrees of freedom. Additionally, estimated regression coefficients can be corrected for overfitting by shrinkage. [3-6]

Given the common dichotomous nature of diagnostic and prognostic outcomes - presence or absence of the outcome or disease - many prediction models are derived with multivariable logistic regression analysis.[2;5] The selection of the predictors is based on the maximization of the log likelihood of the model. While this method provides the optimal fit to the data set at hand, it may result in fitting noise and unstable regression coefficients. Post hoc, one may shrink the regression coefficients by calculating a heuristic shrinkage factor, based on the fitted model chi square corrected for the degrees of freedom that are considered during modelling.[4] Also, bootstrapping methods can be used to estimate a shrinkage factor. [5;8] In both methods a single shrinkage factor is used for all regression coefficients, and the factor is estimated post hoc. Penalised maximum likelihood estimation (PMLE) is a method in which the regression coefficients of a logistic regression model are shrunk individually and immediately during the modelling (likelihood estimation).[9-11] This direct adjustment is the major advantage of PMLE. Instead of maximizing the (conventional) log likelihood, PMLE in fact maximizes the penalized log-likelihood, in which the log-likelihood of the model is adjusted by a penalty factor.

Other methods to derive prediction models for dichotomous outcomes are for example classification and regression trees (CART) and artificial neural networks. However, it has repeatedly been shown that both methods do not produce prediction models that achieve higher predictive performance than models derived by multivariable logistic regression.[12-16] Genetic programming, however, is a more novel and promising search method that may improve the selection of predictors and may lead to models with good predictive accuracy in new patients.[17-20] Genetic programming mimics the natural processes of sexual recombination between two chromosomes (here called crossover) and mutation of DNA on a particular chromosome. First, several models are created from the data. Then, given two selected models, crossover is realized by randomly swapping parts

of the model. Mutation occur by exchanging part of the model with a randomly created substitute. This random process, in addition to the probabilistic nature of the selection step of the models, prevents the search method from converging at a local optimum. Critics state that genetic programming relies more on the data than conventional derivation methods and is therefore more prone to overfitting.

Comparisons have been made between models derived by conventional logistic regression and logistic regression with PMLE[3;11], and between models derived by conventional logistic regression and genetic programming[17], but not between all methods. We therefore derived models that predict the presence or absence of Deep Venous Thrombosis (DVT) with four different derivation methods, namely by logistic regression without shrinkage, logistic regression with shrinkage by bootstrapping techniques, logistic regression with shrinkage by PMLE, and genetic programming (no shrinkage applied). All methods used a derivation set to select the predictors and to test the need for transformations. The accuracy of the models was subsequently tested in cross validations and in new patients of a validation set.

Material and Methods

Data

The clinical example we used concerns the prediction of presence or absence of Deep Venous Thrombosis (DVT). Timely recognition of DVT or ruling out DVT is important because patients with untreated DVT may develop pulmonary embolism whereas unjustified therapy with anticoagulants poses a risk for major bleeding.[21] In primary care, the physician decides on patient history, physical examination and usually the D-dimer assay result, which patients should be referred to the hospital and which can be safely kept under their own surveillance. A diagnostic prediction model can aid physicians in this decision.

We used the data of a cohort of 2086 primary care patients suspected of DVT (415 patients with DVT), included between 2001 and 2005, that has been described previously in the literature.[22;23] After obtaining the patient history, physical examination and the D-dimer result (in 21 total candidate predictors, see Table 1), all patients underwent repeated leg ultrasound as the reference method to determine the true presence or absence of DVT. To compare the effect of the different modelling strategies, we split the cohort into a derivation set (1668 patients, 329 with DVT) to derive the models and a validation set (418 patients, 86 with DVT) to test the models (Table 1).

Testing the differences in the accuracy of models by a single split sample approach can lead to chance findings, since the derivation and validation set are a random sample of the cohort. [5;24] Therefore, we also conducted 100 cross validations in the derivation set to assess the accuracy of the models in different samples. The derivation set was randomly split in ten equal samples. These ten equal samples were combined ten times, in such a way that ten different models were derived on 90% of the data, and tested in the other 10% of the data, while every patient was part of the testing part only once. By splitting the derivation set ten times, 100 cross validations were performed.

Table 1 Distribution of the 21 predictors (and the outcome) in the derivation set and the test set, n (%) unless stated otherwise.

Patient characteristics	Derivation set (n=1668)		Validation set (n=418)	
Age, years[a]	60	(18)	61	(17)
Male gender	603	(36%)	165	(40%)
Oral contraception use	172	(10%)	32	(8%)
Hormonal replacement use	32	(2%)	8	(2%)
Duration of symptoms, days[a]	8	(9)	8	(7)
Absence of a leg trauma	1402	(84%)	353	(84%)
Previous DVT	343	(20%)	78	(19%)
Family history of DVT	341	(20%)	96	(23%)
Presence of a malignancy	95	(6%)	19	(5%)
Immobilisation	217	(13%)	61	(15%)
Recent surgery	198	(12%)	62	(15%)
Swelling whole leg	727	(44%)	205	(49%)
Vein distension	300	(18%)	70	(17%)
Pain in leg	1458	(87%)	347	(83%)
Pain when walking	1370	(92%)	324	(78%)
Oedema in leg	1039	(62%)	271	(65%)
Ill feeling	315	(19%)	85	(20%)
Tender venous system	1213	(73%)	299	(72%)
Pregnancy	38	(2%)	8	(2%)
Log(Calf circumference)[a]	1.07	(0.52)	1 .09	(0.51)
Log(D-dimer level)[a]	6.81	(1.16)	6.89	(1.14)
DVT present	329	(20%)	86	(20%)

[a] Mean (standard deviation)
DVT Deep Venous Thrombosis

Modelling methods

Logistic regression without shrinkage

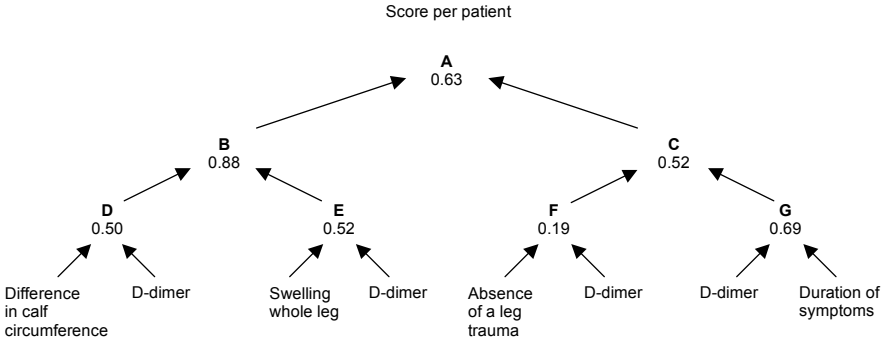
To graphically study the linearity of the continuous predictors, we plotted the association of age, duration of pain, difference in calf circumference and the D-dimer value with the probability of DVT, using restricted cubic spline functions with 3 knots.[5;6] If the graphical plot of the restricted cubic spline showed that the association between the continuous predictors and the outcome was not linear, the predictors were transformed. Subsequently, all 21 candidate predictors were included in a logistic regression model. Backward stepwise selection of the candidate predictors was applied using the likelihood ratio test with a p-value according to Akaike's Information Criterion. Since all candidate predictors had one degree of freedom this corresponds to a p-value > 0.157.

Logistic regression with one overall post hoc shrinkage factor
The same steps were followed as in the previous method. Subsequently, hundred bootstrap samples were drawn from the derivation set. In each bootstrap sample, the modeling process was repeated, including testing for transformations and the backward stepwise selection of the candidate predictors. This results in 100 models that are applied to the original derivation set. The shrinkage factor is equal to the mean of the 100 calibration slopes (see also the section the description of the calibration plots in Accuracy measures). The regression coefficients of the selected predictors of the final model were then multiplied with this shrinkage factor.
Logistic regression with inherent shrinkage by penalised maximum likelihood estimation
As in the conventional logistic regression, the linearity of the continuous predictors was studied and all 21 candidate predictors were included in a logistic regression model. The maximum log likelihood of the full model is adjusted (shrunk) by a penalty factor:
$\text{Log } L - 0.5\ \lambda\ \Sigma(s_i\beta_i)^2$ (Formula 1)
where L is the maximum likelihood of the fitted model, λ a so-called penalty factor, β the estimated regression coefficient for each predictor i in the model, and s_i is a scaling factor for each β_i to make $s_i\ \beta_i$ unitless. [6;9;10] We used a modified AIC for finding the optimal penalty factor. Accordingly, the estimated regression coefficients are directly (i.e., during the model fit) adjusted for overoptimism. This direct adjustment is the major advantage of PMLE. Due to the penalization, the degrees of freedom effectively used in PMLE is lower than the actual number of predictors, reducing the potential for overfitting.[6;9;25]

Parsimonious models can be obtained by estimating a new model with a reduced number of predictors that approximates predictions from the penalised model. The shrinkage that is used in the penalised model is inherited by the model with the reduced number of predictors.[6] To derive the parsimonious model, we use ordinary least squares regression to fit the linear predictor of the penalised model as the outcome and all 21 predictors as covariates. This model necessarily has an R^2 of 1. We use backward stepwise selection to exclude the least important predictors until the R^2 is lower than 0.975. Similar to using a p-value as a threshold to include predictors in the model, the higher the threshold value of the R^2, the more liberal the inclusion and the more predictors will remain in the final model. The regression coefficients of the predictors that remained in this linear regression model are directly the regression coefficients of the reduced penalised prediction model.[6]
Genetic programming
Genetic programming is a search method inspired by the biological model of evolution.[18;26-28] It is an extension of the genetic algorithm first described by Holland[29] and Goldberg[30]. For the present analyses, we used the method of Predictive Analytics Director (Chordiant Software, www.chordiant.com).

In the genetic programming method by Predictive Analytics Director, a prediction model is a mathematical formula that uses a set of candidate predictors as inputs. The building blocks of the formula are mathematical operators, chosen from a library of 20 operators (for example x, x^2, $\sin(\pi*x)$, see also Figure 1). Each operator has two input values and one output value (Figure 1, upper part). The output of the complete formula is a score, which can be related to the predicted probabilities of the outcome under study.

Figure 1 The final model created by genetic programming, presented as a binary tree. The nodes A to G represent the following binary operators, in which the parameters x (left arrow) and y (right arrow) are the inputs of each operator.



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

$B = p(2x - k(2x-1)^3 - 0.5) + (1-p)(2y - k(2y-1)^3 - 0.5)$ and $k = 0.593$

$C = 1 - p\sqrt{(1-x^2)} - (1-p)\sqrt{(1-y^2)}$

$D = p((x-1)^3 + 1) + (1-p)\sin(\pi * y)$

$E = p\sqrt{x} + (1-p)y^2$

$F = px + (1-p)y$

$G = px^2 + (1-p)y^2$

The predicted probabilities can be calculated by $\left(\frac{1}{1+e^{-\text{linear predictor}}}\right)$, in which

linear predictor = -28.79 +37.74* Score -344.41*$(\text{Score} - 0.62)_+^3$+489.40*$(\text{Score} - 0.70)_+^3$-144.99*$(\text{Score} - 0.90)_+^3$
Notation $(x)_+$ means $(x)_+ = x$ if $x > 0$ and 0 otherwise.

In the present study, first a set of 50 different prediction models was randomly created. This set consisted of 50 different mathematical formulas using different predictors. Then in an iterative process, using the data of the derivation set:
1. The ROC area (a measure of discrimination, see the section on Accuracy measures) of each model was estimated.
2. Various models were selected where models with a larger ROC area had a higher probability of being selected. In this way a selection pressure is introduced.
3. New models are created by simulating the natural processes of sexual recombination (here called crossover) between two chromosomes and mutation of DNA on a particular chromosome. In the context of genetic programming a prediction model or binary tree, as selected in step 2, can be compared with a chromosome. Here, crossover and mutation operate on the branches (i.e. parts of the formula having one or more predictors as input, see Figure 1) and nodes of these trees. Given two selected models, crossover is realized by swapping branches between the two trees. The swapped branches are randomly chosen. A mutation occurs by exchanging a node or a branch in a tree with a randomly created substitute, also here the node or branch that is mutated is randomly chosen. This random process, in addition to the probabilistic nature of the selection step (step 2), prevents the search method from converging at a local optimum.

4. These newly created models were moved to the next set of 50 models, which is called the next generation, and upon completing this set the next iteration started
5. The iterative process was terminated after 500 generations when no significant model improvement was observed. The model in the population with the largest ROC area was then selected as the final genetic programming model.
As the mathematical formulas or models all consist of binary operators, they can be represented as a binary tree (see Figure 1, upper part). To limit the amount of overoptimism, the trees were restricted to be no more than 4 levels deep, corresponding to a maximum of 8 predictors.

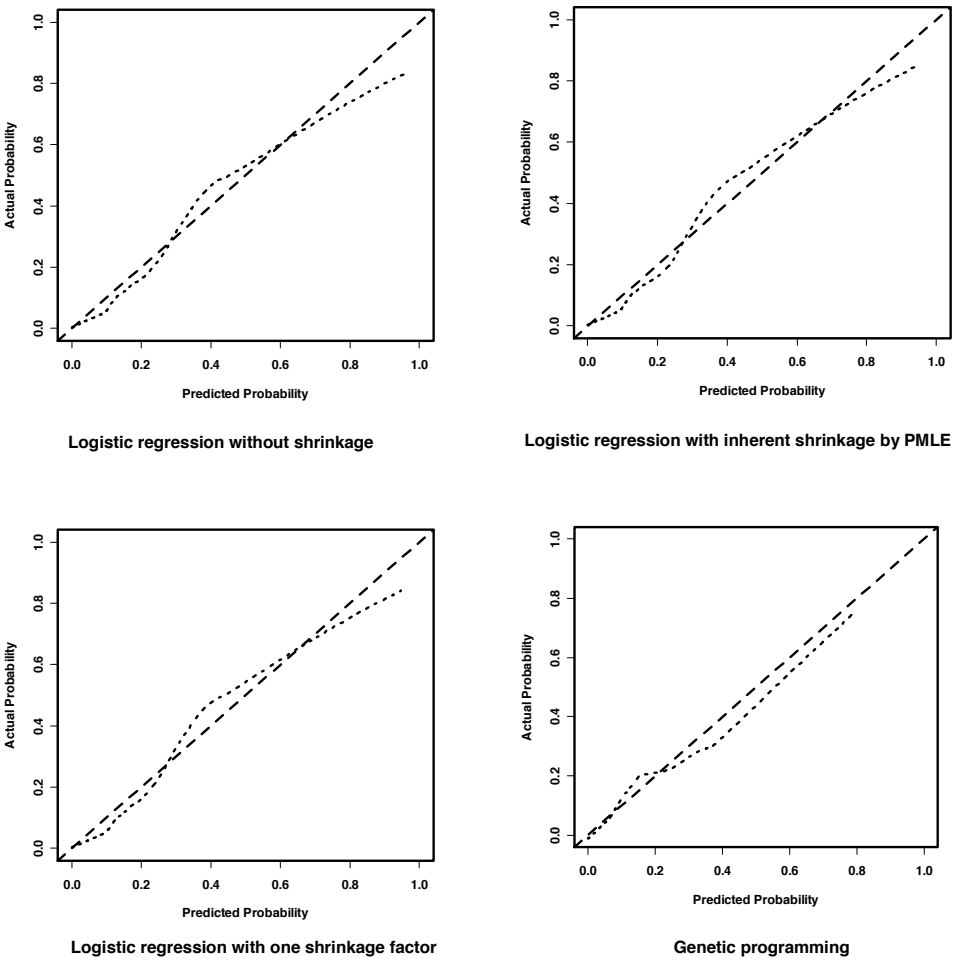
A genetic programming model provides a score for each patient in the derivation set. To transform this score to a probability of DVT presence, we estimated a logistic regression model in the derivation set. The score is the only covariate, modeled with a restricted cubic spline function, as it likely has a non-linear association with the outcome (presence of absence of DVT).[6] Subsequently, this model was used to calculate for each patient the probability of DVT from the corresponding score (see bottom Figure 1). Since shrinkage methods for genetic programming models are less straightforward then in logistic regression models and are not developed yet, the final model could not be adjusted for overfitting.

Accuracy measures

We estimated the accuracy of the models by assessing the discrimination and calibration, in the 100 cross validations of the derivation set and in the validation set. In the cross validation, the median value of the discrimination and calibration were presented. Also, paired t-tests were used to assess whether the differences in the discrimination and calibration between the 100 cross validations were significant.

Discrimination is the ability of the model to distinguish between patients with the outcome and patients without the outcome. It can be quantified with the area under the Receiver Operating Characteristic curve (ROC area), which is equal to the c-statistic for a dichotomous outcome variable.[31] An ROC area ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination).[32] Calibration refers to the agreement between the predicted probabilities and observed frequencies of the outcome. It can be assessed with a calibration plot with the predicted probabilities on the x-axis and the observed frequencies on the y-axis (see Figure 2). The calibration plot shows a calibration line, which can be described with a calibration slope and a calibration intercept.[33;34] These are estimated by fitting the linear predictor of the applied model (original or adjusted) as the only covariate in a logistic regression model. The calibration slope and calibration intercept of a model in new patients are ideally equal to 1 and 0 respectively (perfect calibration). In that situation, the calibration plot lies exactly on the 45˚ line and crosses the y-axis in 0. A slope < 1 indicates optimism (predictions are too extreme), while a slope > 1 indicates that the predictions are not extreme enough. When the slope is not equal to 1, the interpretation of the intercept is difficult. Hence, we estimated the intercept with the slope fixed at 1 (calibration in the large). When this intercept is close to 0, the calibration in the large is good, i.e. the mean predicted probability equals the mean observed proportion.

Figure 2 Calibration plots in the validation set of the models derived by the four derivation methods



Results

Descriptives

Table 1 shows the distribution of the candidate predictors in the derivation and validation set. We found some differences in patient characteristics between the derivation and validation set. The validation set contained more males compared to the derivation set (40% versus 36%). More patients in the validation set experienced a swelling of the whole leg (49% versus 44%), and less patients in the validation set experienced pain when walking (78% versus 92%).

Model derivation

Logistic regression without shrinkage

The difference in calf circumference and the D-dimer value showed a logarithmic association with the presence of DVT and were transformed. The final prediction model included 6 predictors; age, duration of symptoms, absence of a leg trauma, pregnancy, the difference in calf circumference, and the D-dimer value (Table 2). The ROC area in the derivation set was 0.904 (95% Confidence Interval (CI): 0.885 – 0.922).

Logistic regression with one overall post hoc shrinkage factor

By definition the model included the same predictors as the previous model. Bootstrapping resulted in a shrinkage factor of 0.94 that was used to adjust the regression coefficients of the logistic regression model without shrinkage (Table 2). The ROC area was 0.904 (95% CI: 0.885 – 0.922).

Logistic regression with inherent shrinkage by penalised maximum likelihood estimation

The reduced penalised model included 5 predictors; the same predictors as in the previous model though without pregnancy (Table 2). The optimal penalty factor was 3 and the penalised model had 18.93 effective degrees of freedom. The R^2 of the reduced model was 0.98, and the ROC area 0.902 (95% CI: 0.883 – 0.921).

Genetic programming

The final prediction model included the same 5 predictors as the model derived by logistic regression with PMLE, except that age was excluded and swelling of the leg was included (Figure 1). The ROC area was 0.910 (95% CI 0.893 – 0.928).

Table 2 Regression coefficients of the models derived by logistic regression; without shrinkage, with post hoc shrinkage by bootstrapping, and with inherent shrinkage bij PMLE.

	No shrinkage	Bootstrapping	PMLE[a]
Intercept	-15.871	-14.97	-15.562
Age	-0.024	-0.023	-0.021
Duration of symptoms	-0.029	-0.027	-0.027
Absence of leg trauma	-0.688	0.647	0.625
Pregnancy	-1.764	-1.658	-
Log(difference in calf circumference)	1.010	0.949	1.019
Log(D-dimer value)	1.967	1.849	1.892

[a] penalized maximum likelihood estimation

The formula of the above described prediction models take the form:

linear predictor = $ß_0 + ß_1$*predictor$_1 + ß_2$*predictor$_2 + \ldots + ß_n$*predictor$_n$

where $ß_0$ is the intercept and $ß_1$ till $ß_n$ are the regression coefficients of the predictors. The probability of DVT in an individual patient (scale: 0-100%) can be calculated with the formula:

$$\frac{1}{1 + e^{-\text{linear predictor}}}$$

Cross validations

There were only marginal differences in the median discriminative ability of the models in the 100 cross validations (Table 3). The paired t-test showed that there were no significant differences in the ROC areas of the 100 models derived by logistic regression with shrinkage, logistic regression with shrinkage by bootstrapping and logistic regression

Table 3 Comparison of the accuracy measures of the models using the four different derivation methods in hundred cross validations. The accuracy measures (median with 25% and 75% quantiles) include discrimination, expressed by the ROC area, and calibration, expressed by the slope and the intercept | slope =1 (calibration in the large).

Method	Logistic regression	Logistic regression	Logistic regression	Genetic programming
Shrinkage	-	Bootstrap	PMLE [a]	-
ROC area	0.905	0.905	0.903	0.896
(25%-75%)	(0.880 -0.923)	(0.880 -0.923)	(0.880-0.922)	(0.865 – 0.912)
Slope	1.009	1.097	1.067	0.972
(25%-75%)	(0.835 – 1.199)	(0.907 – 1.302)	(0.893 – 1.258)	(0.843 – 1.082)
Intercept | slope = 1	0.002	0.004	0.004	0.003
(95% CI)	(-0.156 – 0.132)	(-0.144 – 0.118)	(-0.104 – 0.152)	(-0.134 – 0.179)

[a] penalized maximum likelihood estimation
25%-75%: 25%-75% quantiles
95% CI: 95% confidence interval

with PMLE. The ROC areas of the 100 models derived by logistic regression without shrinkage or with shrinkage by bootstrapping were significantly different from the ROC areas of the 100 models derived by genetic programming ($p = 0.04$). The ROC areas of the 100 models derived by logistic regression with PMLE were significantly different from the ROC areas of the 100 models derived by genetic programming ($p = 0.05$).

All methods led to models with a median slope close to 1. The median slope of the models derived by genetic programming was smaller than 1, indicating overfitting. The paired t-tests showed a few significant differences between the slope of the 100 models in the cross validations, notably between the models derived by logistic regression without shrinkage and the models derived by logistic regression with bootstrapping ($p=0.01$), between the models derived by logistic regression with bootstrapping and the models derived by genetic programming ($p<0.001$), and between the models derived by logistic regression with PMLE and the models derived by genetic programming ($p<0.001$). The median intercept (given slope =1) was close to 0 for all models and was not significantly different between the 100 models derived by the different methods.

Model validation

We found only marginal differences in the discriminative ability of the four models in the validation set (Table 4). The discriminative ability of all models was slightly higher in the validation set compared to the derivation set. All methods led to models with a slope close to 1 (Table 4 and Figure 2). The slopes of the models derived by logistic regression without shrinkage and genetic programming were smaller than 1, indicating overfitting (Table 4). The intercept given that the slope was equal to 1 was close to 0 for all models (Table 4 and Figure 2). Logistic regression with PMLE led to the best calibration in the large (intercept = -0.087) and genetic programming to the worst (intercept = -0.160), although differences were small

Table 4 Comparison of the accuracy measures of the models using the four different derivation methods in the test set. The accuracy measures include discrimination, expressed by the ROC area, and calibration, expressed by the slope and the intercept | slope =1 (calibration in the large).

Method	Logistic regression	Logistic regression	Logistic regression	Genetic programming
Shrinkage	**-**	**Bootstrap**	**PMLE[a]**	**-**
ROC area	0.906	0.906	0.907	0.912
(95% CI)	(0.872 -0.941)	(0.872 -0.941)	(0.873-0.941)	(0.882-0.943)
Slope	0.961	1.024	1.027	0.982
(95% CI)	(0.756 – 1.169)	(0.804 – 1.244)	(0.808 – 1.247)	(0.788 – 1.176)
Intercept | slope = 1	-0.142	-0.105	-0.087	-0.160
(95% CI)	(-0.455 – 0.171)	(-0.411 – 0.201)	(-0.394 – 0.220)	(-0.480 – 0.161)

[a] penalized maximum likelihood estimation
95% CI: 95% confidence interval

Discussion

We derived a clinical prediction model to predict the presence or absence of DVT with four different derivation methods: conventional logistic regression without shrinkage, logistic regression with one overall shrinkage factor (estimated from bootstrapping techniques), logistic regression with shrinkage by PMLE, and genetic programming. The accuracy of the methods was tested in cross validations and in new patients (external validation). There were only marginal differences in the median discriminative ability of the models in the 100 cross validations. All methods led to models with a median slope close to 1, and a calibration intercept (given slope = 1) close to 0 for all models in the 100 cross validations. Similarly, the accuracy measures of the four models in the validation sets were only slightly different, and the 95% confidence intervals of the outcome measures largely overlapped.

We expected that the model derived by logistic regression with PMLE would show better discriminative ability (due to less overfitting) in the validation set than the other models derived by logistic regression, as PMLE adjusts the individual regression coefficients during the estimation process. Hence, PMLE is theoretically capable of improving the discriminative ability of a prediction model when applied in new patients. This was not shown in our data. Also, we expected that the slopes of the models derived by logistic regression without shrinkage and genetic programming would be smaller than 1, indicating overfitting. This was indeed shown in the validation set, but only partly in the cross validations, where only the models derived by genetic programming resulted in a median slope smaller than 1.

We can think of several reasons why we found no major differences in the accuracy (both discrimination and calibration) of the models derived by the four methods. First, overfitting of models often occurs in relatively small datasets. [3-6] Our derivation set consisted of data of 1668 patients, of which 329 patients (20%) had DVT. According to the rule of thumb that at least ten events are needed to consider one candidate predictor (with

one degree of freedom), this implies that 33 candidate predictors could have been considered. We considered only 21 candidate predictors, using eight extra degrees of freedom for assessing the linearity of the four continuous predictors with restricted cubic splines. Hence, with in total 29 degrees we still conformed to that rule of thumb, reducing the risk of overfitting. Second, one of the strengths of genetic programming is the flexibility in the modelling of the continuous predictors. However, in our dataset only 4 of the candidate predictors were continuous predictors, and the other 17 predictors were dichotomous predictors. In a dataset with relatively more continuous predictors the discriminative ability of the model derived by genetic programming may have been higher. Note that the use of restricted cubic spline functions also increases the flexibility of modelling continuous predictors in conventional logistic regression. Third, the effectively used number of degrees of freedom in PMLE was only slightly lower than the actual number of predictors (18.93 versus 21). The ability of PMLE to reduce the potential for overfitting could have been more obvious when interaction terms were included in our prediction model. The effectively used number of degrees of freedom would then be considerably lower than the actual number of predictors, with a better accuracy of the PMLE model than the model derived by conventional logistic regression in new patients. As we had no clinical reasons to include interaction terms in our dataset, we did not include these.

All four methods have their own opportunities and pitfalls, depending on the characteristics of the data and situation at hand. First, when conducting logistic regression, one should preferably apply shrinkage, especially in small datasets.[5;6;35;36] Although the discriminative ability of a model will not be changed, in general the calibration of the model in future patients will be improved when a shrinkage method is applied. Second, when many interaction terms are considered, PMLE should be preferred over shrinkage by one overall shrinkage factor, especially when the dataset is (relatively) small. Third, one of the strengths of genetic programming is that it is so-called self-learning software. Hence, when the calibration decreases due to discrepancies that arise between the predicted probabilities and the observed frequencies a model is applied to new patients, the model can be re-calibrated with the information of the new patients. As a result, the mean predicted probabilities remain equal the mean observed frequencies. Also models derived by logistic regression can be re-calibrated or adjusted to the circumstances of new patients, by so-called updating methods. [37-40] Yet, genetic programming can automatically re-calibrate the prediction model, while updating a logistic regression model requires interference of researchers. The increasing use of electronic patient records (EPR) in medical care in stead of conventional paper files facilitates re-calibration or updating of prediction models. Once a prediction model is incorporated in the EPR, information of new patients is constantly added to the EPR facilitating regular updating of the models. Fourth, genetic programming is highly flexible in the transformations of continuous predictors. Especially in datasets with many continuous predictors, genetic programming can provide insight in unconventional associations between the predictor and the outcome. Note that the use of restricted cubic spline functions and fractional polynomials[41-43] in logistic regression also increases the flexibility of modelling continuous predictors. Note that more degrees of freedom are needed to permit this flexible modelling, potentially increasing the risk of overfitting. Fifth, the models derived by genetic programming can not be shrunk according to standard methodology, potentially leading to overfitted prediction models. This is particularly precarious when a prediction model is used to inform patients about their

absolute probability of a disease or event. Also when the predicted probability is used to start or withhold treatment, the correct estimation of the absolute probability is a precarious issue. Therefore, these shrinkage methods should be explored in the future.

In conclusion, in this empirical example the predictive accuracy of the models derived by logistic regression, logistic regression with shrinkage by bootstrapping techniques, logistic regression with shrinkage by PMLE, and genetic programming was not different in our dataset, and the 95% confidence intervals of the accuracy measures largely overlapped. The choice between these derivation methods should be based on the characteristics of the data and situation at hand.

References

1 Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993; 118(3):201-210.

2 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997; 277(6):488-494.

3 Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000; 19(8):1059-1079.

4 Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990; 9(11):1303-1325.

5 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(4):361-387.

6 Harrell FE, Jr. Regression modelling strategies. Springer-Verlag, New York; 2001.

7 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49(12):1373-1379.

8 Efron B. Censored Data and the Bootstrap. Journal of the American Stat Association 1981; 76(374):312-319.

9 Gray RJ. Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 1997; 87:942-951.

10 Verweij PJ, van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med 1994; 13(23-24):2427-2436.

11 Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 2004; 57(12):1262-1270.

12 Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Investig Med 1995; 43(5):468-476.

13 Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. 2001. J Clin Epidemiology. Ref Type: Generic

14 Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. 1996. J Clin Epidemiol. Ref Type: Generic

15 Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression

methods to diagnose myocardial infarction. Medinfo 1998; 9 Pt 1:493-7.:493-497.
16 Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998; 17(21):2501-2508.
17 Biesheuvel CJ, Siccama I, Grobbee DE, Moons KG. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. J Clin Epidemiol 2004; 57(6):551-560.
18 Forrest S. Genetic algorithms: principles of natural selection applied to computation. Science 1993; 261(5123):872-878.
19 Podbregar M, Kovacic M, Podbregar-Mars A, Brezocnik M. Predicting defibrillation success by 'genetic' programming in patients with out-of-hospital cardiac arrest. Resuscitation 2003; 57(2):153-159.
20 Tsakonas A, Dounias G, Jantzen J, Axer H, Bjerregaard B, von Keyserlingk DG. Evolving rule-based systems in two medical domains using genetic programming. Artif Intell Med 2004; 32(3):195-216.
21 Hirsh J, Hoak J. Management of deep vein thrombosis and pulmonary embolism. A statement for healthcare professionals. Council on Thrombosis (in consultation with the Council on Cardiovascular Radiology), American Heart Association. Circulation 1996; 93(12):2212-2245.
22 Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005; 94(1):200-205.
23 Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006; 55(7):613-618.
24 Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall; 1993.
25 Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Stat Med 2002; 21(24):3803-3822.
26 Barrett J, Kostadinova A, Raga JA. Mining parasite data using genetic programming. Trends Parasitol 2005; 21(5):207-209.
27 Koza J.R. Genetic Programming III. Cambridge, MA: MIT Press; 1999.
28 Poli R, McPhee NF. General schema theory for genetic programming with subtree-swapping crossover: part I. Evol Comput 2003; 11(1):53-66.
29 Holland JH. Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press; 1975.
30 Goldberg DE. Genetic Algorithms in Search Optimization and Machine Learning. Addison Wesley Publishing Company; 1989.
31 Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3(2):143-152.
32 Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148(3):839-843.
33 Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993; 13(1):49-58.
34 Cox DR. Two further applications of a model for binary regression. Biometrika 1958; 45:562-565.
35 Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54(8):774-781.

36 Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003; 56(5):441-447.
37 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23(16):2567-2586.
38 Hosmer D.W., Lemeshow S. Applied logistic regression. New York: John Wiley and Sons, Inc.; 1989.
39 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999; 99(16):2098-2104.
40 van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995; 14(18):1999-2008.
41 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999; 28(5):964-974.
42 Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. Stat Med 2003; 22(4):639-659.
43 Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 2004; 23(16): 2509-2525.

Chapter 3

Validation and updating

Chapter 3.1

Validation, Updating and Impact of Clinical Prediction Rules: a review

Abstract
Prediction rules combine multiple predictors, such as patient characteristics, test results, and other disease characteristics, to estimate the probability that a certain outcome is present (diagnosis) or will occur (prognosis) in individual patients. When applied to new patients, however, the accuracy of prediction rules is generally lower compared to the accuracy in the data on which the prediction rule was developed. Therefore, the accuracy of prediction rules should be tested or validated in new patients before implementation in daily care. Understandably, prediction rules that fail to maintain their accuracy in new patients should not be applied in clinical practice. However, these rules should not directly be rejected, because adjusting the prediction rules (updating) has many advantages compared to developing new rules. Prediction rules which preserve their accuracy when applied to new patients may subsequently be subjected to an impact study, to demonstrate whether they change physicians decisions, improve patient outcome and/or reduce costs. This paper provides an overview of existing methods for validation of prediction rules, solutions to update a rule in case of disappointing accuracy in a validation study, methods for impact studies and we discuss the implementation of prediction rules.

Introduction

Prediction rules or prediction models, often also referred to as decision rules or risk scores, combine multiple predictors, such as patient characteristics, test results, and other disease characteristics, to estimate the probability that a certain outcome is present (diagnosis) in an individual or will occur (prognosis). They intend to aid the physician in making medical decisions and in informing patients. Table 1 shows an example of a prediction rule.

Table 1 Prediction rule for estimating the probability of neurological sequelae or death after bacterial meningitis[26]. The relative weights of the predictors are expressed with regression coefficients and with simplified scores.

	Regression coefficient[a]	Score[b]
Male gender	1.48	2
Atypical convulsions	2.27	3
Body temperature -35oC	-0.75	-1
Pathogen		
S. Pneumoniae	3.12	4
Neisseria meningitidis	1.48	2
Intercept / constant	25.2	5

Score	< 2.5	2.5-4.5	5.0-5.5	> 5.5	Overall
Probability	0/25 (0%)	2/78 (2.6%)	6/33 (18.2%)	15/34 (44.1%)	23/170 (13.5%)

[a] The probability for each patient can be calculated as

$$\log\left(\frac{\text{risk of outcome}}{1-\text{risk of outcome}}\right) = 25.2 + 1.48*\text{male gender} + 2.27*\text{atypical convulsions} - 0.75*(\text{body temperature} - 35^{o}\text{C}) + 3.12*\text{S.Pneumoniae} + 1.48*\text{N. Meningitidis}$$

[b] The scores were derived by dividing the regression coefficients of the included predictors by the smallest regression coefficient and then rounding them to the nearest integer. For each patient, a sumscore can be calculated by adding the scores that correspond to the characteristics of the patient. The total sumscore are related to the individual probability as shown in the lower part of Table 1.

In multivariable prediction research, three phases may be distinguished; 1. development of the prediction rule; 2. external validation of the prediction rule (further referred to as 'validation'), i.e. testing the rule's accuracy and thus generalisability in data that was not used for the development of the rule, and subsequent updating if validity is disappointing; and 3. studying the clinical impact of a rule on physician's behaviour and patient outcome (Table 2)[1-5]. A quick Medline-search using a suggested search strategy[6] demonstrated that the number of scientific articles discussing prediction rules has more than doubled in the last decade; 6,744 published articles in 1995 compared to 15,662 in 2005. A striking fact is that this mainly includes papers concerning the development of prediction rules. A relatively small number regards the validation of rules and there are hardly any publications showing whether an implemented rule has impact on physicians behaviour or patient outcome[3, 4].

Table 2 Consecutive phases in multivariable prediction research.

Phase	Short description
1 Development	Development of a multivariable prediction rule, including identification of important predictors, assigning the relative weights to each predictor, estimating the rule's predictive accuracy, estimating the rule's potential for optimism using so-called internal validation techniques, and -if necessary- adjusting the rule for overfitting.
2 Validation and updating	Testing the accuracy of the prediction rule in patients that were not included in the development study. Temporal, geographical, and domain validation can be distinguished. If necessary, the prediction rule can be updated, by combining the information captured in the rule (development study) and the data of the new patients (validation study).
3 Impact	Determining whether a (validated) prediction rule is used by physicians, changes therapeutic decisions, improves patient outcome or will reduce costs.

Lack of validation and impact studies is unfortunate, because accurate predictions -commonly expressed in good calibration (agreement between predicted probabilities and observed outcome frequencies) and good discrimination (ability to distinguish between patients with and without the outcome)- in the patients that were used to develop a rule are no guarantee for good predictions in new patients, let alone for their use by physicians[1, 3, 4, 7, 8]. In fact, most prediction rules commonly show a reduced accuracy when validated in new patients[1, 3, 4, 7, 8]. There may be two main reasons for this: 1. the rule was inadequately developed; or 2. there were (major) differences between the derivation and validation population.

Many guidelines regarding the development of prediction rules have been published, including the number of potential predictors in relation to the number of patients, methods for predictor selection, how to assign the weights per predictor, how to shrink the regression coefficients to prevent overfitting, and how to estimate the rule's potential for optimism using so-called internal validation techniques such as bootstrapping[1, 2, 7-14].

Compared to the literature on the development of prediction rules, the methodology for validation and studying the impact of prediction rules is underappreciated[1, 4, 8]. This paper provides a short overview of the types of validation studies, of possible methods to improve or update a previously developed rule in case of disappointing accuracy in a validation study, and of important aspects of impact studies and implementation of prediction rules. We focus on prediction rules developed by logistic regression analysis, but the issues largely apply to prediction rules developed by other methods such as Cox proportional hazard analysis or neural networks.

Examples of disappointing accuracy of prediction rules

Even when internal validation techniques are applied to correct for overfitting and optimism, the accuracy of prediction rules can be substantially lower in new patients compared to the accuracy found in the patients of the development population. For example, the generalisability of an internally validated prediction rule for diagnosing a serious

bacterial infection in children presenting with fever without apparent source was disappointing[15]. In the development study, the area under the receiver operating characteristic curve (ROC area) -after adjustment for optimism- was 0.76 (95% confidence interval [CI]: 0.66–0.86). However, when applied to new patients obtained from another hospital in a later period using the same in- and exclusion criteria, the ROC area dropped to 0.57 (95% CI: 0.47-0.67). The authors concluded that this could partly have been caused by flaws in the development of the rule, notably too few patients in relation to the number of predictors tested, but also that internal validation and correction for optimism do not always prevent a decreased accuracy in future patients[15].

Another example of poor generalisability regards the European System for Cardiac Operative Risk Evaluation (EuroSCORE), a prediction rule that was developed in 128 centres in eight European states to predict 30-day mortality in patients who underwent cardiac surgery[16, 17]. Validation studies showed good results in European, North American and Japanese populations[16, 18-23]. Yap et al.[24] tested the generalisability of the EuroSCORE in 8331 cardiac surgery patients from six Australian institutions and found that predictions were poorly calibrated for Australian patients. According to the authors, reasons for this finding are unclear and likely to be multi-factorial, such as different health care system, different indications for cardiac surgery, and different prevelance of co-morbid conditions in Australia compared to Europe. Also, the prediction rule could be 'out of date'[24, 25]. The rule was developed with data of patients who were operated more than 10 years prior to the patients in the Australian validation study. The surgical procedure has indeed changed over time, potentially leading to different outcomes[24, 25]. This change was not reflected by the other validation studies[16, 18-23].

Common differences between a development and a validation population

As described, disappointing generalisability can be explained by differences in the development and validation population. We may largely identify three possible differences. First, the definitions of predictors and the outcome variable, and the measurement methods may be different[1, 2, 8, 11]. Prediction rules that contain unclear defined predictors or predictors which measurement or interpretation is liable to subjectivity are likely to show a reduced predictive strength when applied to new patients. For example, the prediction rule of Table 1 contains the rather objective predictor gender, but also the presence of 'atypical convulsions'[26]. The latter may be defined differently by physicians, which may compromise the generalisability of the rule. It is advised to determine the interobserver variability of potential predictors, and to include only those predictors in the final prediction rule that show good reliability[2, 5]. Improvement in measurement techniques for predictors may also affect the predictive strength of a predictor. For example, the Magnetic Resonance Imaging (MRI) technique is developing rapidly over time, which results in improved image quality. Consequently, the diagnostic or prognostic information of MRI probably also improves over time and influences the accuracy of prediction rules that include MRI information.

Second, the group of patients used for the development of a prediction rule may be different from the group of patients used for validation. This is also called difference in 'case-mix'[1, 27]. For example, differences in indication for cardiac surgery and differences in co-morbidity were considered as one of the causes of the poor calibration of the (prognostic) EuroSCORE in Australian patients. Both discrimination and calibration of a rule can be affected by differences in case-mix. For example, a validation population may only include elderly (e.g. defined as age ≥ 65 years), while in the development population individuals' age ranged from 18-85 years. If age is a predictor in the rule, then

discrimination between presence or absence of the outcome in the more homogeneous validation population is more difficult than in the more heterogeneous development population. Further, a validation population may e.g. contain relatively more males than the development population. If male gender increases the probability of the outcome but gender was not included in the rule (missed predictor), then the predicted probabilities by the rule will be underestimated in the validation population (reduced calibration)[27].

Third, validation studies commonly include fewer individuals than development studies. Accordingly, both populations may *seem* different, which is notably due to random variation[12, 28]. The required size of a validation study depends on the hypotheses tested. For prediction rules that predict dichotomous outcomes, it has been suggested that the validation sample should contain at least 100 events and 100 non-events to detect substantial changes in accuracy with 80% power, e.g. a 0.1 change in c-statistic[29].

Type of validation studies

It has repeatedly been suggested that a validation study should consist of an adequate sample of 'different but related patients' compared to the development study population [1, 4, 8]. Relatedness is at least defined as 'patients suspected of the same disease' for a diagnostic rule, and for a prognostic rule as 'patients at risk of the same event'.

In hierarchy of increasingly stringent validation strategies, we largely distinguish temporal, geographical, and domain validation[1, 3, 4, 8]. In general, the potential for differences between the development and validation population is smallest in a temporal validation study, and largest in a domain validation study (Table 3). Consequently, confirmative results in a domain validation study are considered to provide the strongest evidence that the prediction rule can be generalised to new patients, while the generalisability of a prediction rule which has shown confirmative results in a temporal validation study may still be restricted.

Temporal validation

Temporal validation tests the generalisability of a prediction rule 'over time'. In a temporal validation study, the prediction rule is typically tested by the same physicians or investigators as in the development study, in the same institution(s), and in similar patients e.g. using the same eligibility criteria resulting in small variation in case-mix (Table 3)[3, 4, 8]. Hence, this type of validation is usually successful for thoughtfully developed prediction rules. However, improvements in medical techniques may still affect the predictive accuracy of a prediction rule, as in the example of the EuroSCORE. Confirmative results in (multiple) temporal validation studies indicate that clinicians may cautiously use the prediction rule in their future patients who are similar to the development and validation population. But validation in varied study sites is still necessary before the rule can be implemented in other (geographical) locations or other patient domains (see below)[1, 4, 8].

Geographical validation

Geographical validation studies typically test the generalisability of a prediction rule in a patient population that is similarly defined as the development population (as is the case in temporal validation studies), though in hospitals or institutions of other geographical areas[1, 8]. 'Other geographical areas' can be within one country, across similar countries (e.g. western to western or non-western to non-western countries), and across non-similar countries (e.g. western to non-western or vice versa). Understandably, the less similar the development and validation locations are, the more potential for differences in 'interpretation of predictors and outcome', 'measurements used' and 'case-mix', and thus the more potential for disappointing generalisability (Table 3).

Physicians participating in a geographical validation study may be less experienced using the prediction rule, which may influence the accuracy of the rule, due to predictors sensitive to subjectivity. Further, measurement of predictors and the outcome can be performed with different methods than in the development study, also potentially affecting the rule's accuracy. For example, prediction rules to safely exclude the diagnosis deep vein thrombosis (DVT) contain items from patient history, physical examination, and the result of a D-dimer test[30, 31]. Many different D-dimer assays are available, each with different diagnostic accuracy measures. Sensitivities of D-dimer tests for diagnosing DVT may vary between 48% and 100%, and specificities from 5% to 100%[32]. This variation in the accuracy of different D-dimer tests will obviously be reflected in the accuracy of the diagnostic rules that include D-dimer tests. Further, if different cut-off values are used to dichotomise a continuous variable to define a positive versus negative result, the predictive accuracy of the same variable may be different across studies. For instance, if an abnormal D-dimer concentration has been defined as higher than 500 ng/ml in the development sample and as higher than 1,000 ng/ml in the validation sample, the rule will likely perform differently in the validation sample.

Case-mix differences in a geographical validation study can be subtle. For example, a rule containing 'antibiotic use in previous month' as a predictor to estimate the probability of 30-day hospitalisation or death from lower respiratory tract infection in elderly patients[33] may show decreased accuracy in another country, not because the predictor was not well defined or liable to subjectivity, but because the indication for prescribing antibiotics, and thus the characteristics of antibiotic receivers, may vary between countries.

Domain validation

Perhaps the broadest form of validation is to test the generalisability of a prediction rule across different domains, such as patients from a different setting (primary, secondary or tertiary care), inpatients versus outpatients, patients of different age categories (e.g. adults versus adolescents or children), of a different gender, and perhaps from a different type of hospital (academic versus general hospital). Obviously, the case-mix of a patient population of a new domain will differ from the development population (Table 3), which is usually reflected in differences in the distribution of the predictor values and in the ranges of predictor values.

For example, the case-mix of primary care and secondary care patients is often clearly different. Primary care physicians always selectively refer patients to specialists. These referred patients commonly have relatively more severe signs or symptoms, or have a relatively more developed disease stage[8, 34]. Consequently, secondary care patients commonly have a narrower range of the (more severe) predictor values than primary care patients. To some extent, a secondary care population can be considered as a subdomain of the primary care population. Hence, in contrast, validating a prediction rule developed from secondary care in a more heterogeneous primary care population actually concerns the estimation of the rule's ability for extrapolation. Extrapolation of prediction rules developed in secondary care to primary care patients often results in a decreased accuracy[8, 34]. For example, it has been shown that a prediction rule for safely excluding the diagnosis DVT developed in secondary care[31, 35] showed disappointing accuracy in primary care: 0.9% of the secondary care patients had DVT while the diagnosis was ruled out according the rule, whereas this proportion was increased to 2.9% in primary care patients[36].

We note, however, that certain inclusion criteria in development studies may have been chosen for practical reasons only, and may not compromise the generalisation of a prediction rule derived from such studies. For example, the Ottawa ankle rule for safely excluding fractures without additional *x* ray testing[37] was developed on patients aged 18 years or older. One could question whether this rule -which does not include age as a predictor- is accurate when applied to children or adolescents. If the relative weights (odds ratios) of the predictors in the rule are independent of age, extrapolation of the rule to adolescents can be as successful as in the development population. A recent review concerning the extrapolation of the Ottawa ankle rule to children or adolescents indeed concluded that 'a small' percentage (1.4%) of patients that are excluded from receiving *x* ray evaluation based on the Ottawa ankle rule will actually have a fracture[38], compared to 0% in the development study (in the development study, the score threshold was specifically chosen to achieve a 100% negative predictive value).

Table 3 Potential differences between the development and validation population and the influence on generalisability of the prediction rule: - weak; +/- possible; + probable; ++ likely

	Temporal validation	**Geographical validation**	**Domain validation**
	To test the generalisability of a prediction rule 'over time' in similar patients as in the development study (same in- and exclusion criteria) in the same hospitals or institutions.	To test the generalisability of a prediction rule in similar patients as in the development study (same in- and exclusion criteria) in hospitals or institutions of another geographical area.	To test the generalisability of a prediction rule across different domains, which may contain other patient (sub)groups.
Differences in:			
Interpretation of predictors and outcome	-	+	+
	Often same physicians as in the development study who are thus experienced in obtaining the predictors and outcome; differences in interpretation of predictors is less unlikely	Other physicians as in the development study, who may define subjective predictors (and the outcome) differently; differences in interpretation of predictors may occur	Other physicians as in the development study, who may define subjective predictors (and the outcome) differently; differences in interpretation of predictors may occur
Used measurements for predictors and outcome	+/-	+	+
	Same measurements used, unless measurements have been replaced (time-related)	Other measurements may be used ('institution-dependent'), plus potential for time-related changes in measurements	Other measurements may be used ('institution-dependent'), plus potential for time-related changes in measurements
Case-mix	-	+/-	++
	Patient populations are similar; (random) variation due to a commonly small sample size of the validation population is possible	(Subtle) differences in case-mix possible, beyond possible (random) variation due to a commonly small sample size of the validation population	Differences in case-mix are (very) likely, beyond possible (random) variation due to a commonly small sample size of the validation population

Updating prediction rules

When a validation study shows disappointing results, researchers are often tempted to reject the rule and directly pursue to develop new rules with the data of the validation population only. However, while the original prediction rules usually have been developed with large datasets, validation studies are frequently conducted with much smaller patient samples. The redeveloped rules are thus also based on smaller samples. Furthermore, it would lead to many prediction rules for the same outcome, obviously creating impractical situations as physicians have to decide on which rule to use. For example, there are over 60 published rules to predict outcome after breast cancer[39]. Moreover, when every new patient sample would lead to a new prediction rule, prior information that is captured in previous studies and prediction rules would be neglected. This is counterintuitive to the intention that scientific inferences should be based on data of as many patients as possible. This principle of using prior knowledge from previous studies has been recognized and utilized in etiologic and intervention research, for example in the realm of (cumulative) meta-analyses.

A logical alternative to re-developing prediction rules in each new patient sample, is to update existing prediction rules with the data of the new patients in the validation study. As a result, updated rules combine the prior information that is captured in the original rules with the information of the new patients of the validation population[40-43]. Hence, updated rules are adjusted to the characteristics of the new patients, and likely show improved generalisibility.

Several updating methods have been proposed in the literature[40-43]. The methods vary in extensiveness, which is reflected by the number of parameters that is adjusted or re-estimated (Table 4). We will briefly describe these methods and refer to the literature for a more profound description[40-43].

Table 4 Updating methods for prediction rules.

No.	Updating method	Reason for updating
0	No adjustment (the original prediction rule)	-
1	Adjustment of the intercept	Difference in outcome incidence
2	Adjustment of the regression coefficients of the predictors (by one adjustment factor) and of the intercept	Regression coefficients of the original rule are overfitted
3	Method 2 + extra adjustment of regression coefficients for predictors with a different strength in the validation population compared to the development population	As in method 2, and the strength (regression coefficient) of one or more predictors may be different in the validation population
4	Method 2 + stepwise selection of additional predictors	As in method 2, and one or more potential predictors were not included in the original rule
5	Re-estimation of all regression coefficients, using the data of the validation population	The strength of all predictors may be different in the validation population
6	Model 5 + stepwise selection of additional predictors	As in method 5, and one or more potential predictors were not included in the original rule

In many situations, as described before, differences in outcome incidence are found between the development data and the validation data. For example, in a primary care setting one can validate a secondary care rule that predicts the presence or absence of DVT. Due to the higher prevalence of DVT in secondary care[34], the calibration of the rule in primary care patients may be poor as a result of systematically too high predicted probabilities. By adjusting only the intercept of the original prediction rule for the patients in the primary care setting, the poor calibration can be improved[42, 43]. This method is by far the simplest updating method as only one parameter of the original rule, i.e. the intercept, is adjusted (Table 4, method 1).

Another updating method is called 'logistic recalibration' and can be used when the regression coefficients (relative weights that represent the predictive strength) of the predictors in the prediction rule are overfitted in the development study[42, 43]. This typically results when too many predictors were considered in a too small dataset[7, 13]. In the lower range, the predicted probabilities in the new patients are usually too low, while in the higher range they are too high. When overfitting was not adequately prevented or adjusted during development of the rule, all regression coefficients can still be adjusted with a single correction factor that is easily estimated from the data of the new patients in the validation set (Table 4, method 2).

Although calibration can indeed be improved by these first two methods, discrimination (ROC area) will remain unchanged, as the relative ranking of the predicted probabilities remain the same. To improve the discrimination of a rule in new patients, more rigorous adjustments need to be made to the prediction rule, also called model revisions. We will briefly explain four revision methods that can also improve the discrimination of a prediction rule when a rule is validated in new patients.

First, the strength of one or more predictors can be different in the validation population compared to the development population, while the relative sizes of the other regression coefficients to each other are correct. The regression coefficients that differ can be re-estimated from the validation data (Table 4, method 3). For example, we discussed a prediction rule with the antibiotic use as a predictor (section 2.2.2). When this rule is applied in a population or setting with a different antibiotic prescription strategy, the strength of this particular predictor may be different, while the strength of the other predictors in the rule not necessarily changes.

Also, when potential predictors that may have predictive value were not included in the original rule, one can test whether these have added predictive value in the validation data (Table 4, method 4). For example, when a prediction rule is validated over time, and a new test has become available, the new test may have added predictive value in the rule.

Finally, when the previous described updating methods can not improve the accuracy of the rule, and the strengths of all predictor are expected to be different in the new patients, the intercept and the regression coefficients of all predictors can be re-estimated with the data of the new patients (Table 4, method 5). If necessary, additional predictors can be considered as well (Table 4, method 6). These two methods are the most rigorous updating methods, as the intercept, regression coefficients and, possibly also additional predictors, are all re-estimated from the validation set. Both methods will probably be most applicable to domain validation, as these are typical situations in which the strength of predictors may differ between the two populations. Note that a disadvantage of these

rigorous updating methods is that the rule is redeveloped on the data of the validation set only and that the prior information in the original rule is neglected, as we discussed above.

With all above described methods, the updated rules are adjusted to the circumstances of the validation population. However, we recommend that updated prediction rules, just like newly developed rules, still need to be tested on their generalisability and impact before they can be applied in daily practice. Note that for all updating methods, data of the new patients is needed. When this data is not available, but one knows the incidence of the outcome and the mean value of the predictors in the new population, the rule can be adjusted by a simple adjustment of the prediction rule[44, 45].

Impact analysis

To ascertain whether a validated prediction rule will actually be used by physicians, will change or direct physicians' decisions, and will improve patient outcomes or reduces costs, an impact study or impact analysis should be performed[3, 4]. In the ideal design of an impact study, physicians or care units are randomized to either the index group – which is 'exposed' to the use of the prediction rule – or to the control group using 'care or clinical judgment as usual'[3]. Randomization of patients instead of physicians -such that a physician randomly uses the prediction rule or applies 'usual care'- is not advised. Learning effects will lead to a reduced contrast between the two study groups, resulting in a diluted measured impact of the rule. Moreover, randomising centres (requiring a multi-centre study) instead of physicians within a single centre may prevent the risk of contamination – i.e. exchange of experiences and information by physicians between the two study groups – also leading to reduced contrast and dilution of the rule's effect. Patients are followed up to determine the impact of the prediction rule on patient outcome and on cost-effectiveness. Follow-up is not required for studying the influence of a prediction rule on decision making behavior: a randomized study then suffices. An alternative design to determine the impact of a prediction rule is a before-after study within the same physicians or care units[46], although temporal changes may compromise the validity of this design.

Although an impact analysis is the method par excellence to study the real effect of a prediction rule in practice, only a limited number of impact analyses have been performed[4]. One of these few studies regarded the impact of a prediction rule aimed at improving the effectiveness of treatment of patients presenting with community-acquired pneumonia to the emergency department, measured by the health-related quality of life and the number of bed days per patient[47]. Hospitals were randomly assigned to implementation of the prediction rule or conventional care, by using a computer that stratified for type of institution (teaching or community hospital) and average historical length of stay. Physicians were instructed to use the prediction rule as a guide only; the rule did not supersede clinical judgment. An educational plan was designed to reinforce compliance with the use of the prediction rule. The authors concluded that the prediction rule did not improve the health-related quality of life of the patients but reduced the number of bed days per patient managed. This effect was never revealed if no impact study had been conducted.

Implementation of prediction rules

When a rule has frequently been proven to be accurate in diverse populations, the more likely it is that the prediction rule can be successfully applied in practice[1, 4, 8]. Yet, there are still reasons why the rule is not as successful in daily practice.

First, physicians may *feel* that their often implicit estimation of a particular predicted probability is at least as good as the probability calculated with a prediction rule, and may therefore not use or follow the rule's predictions[3]. Or, the physicians' estimation of a probability may even have proven to outperform the discrimination of a prediction rule. In a recent review, Sinuff *et al.* compared the discrimination of physicians' estimations with prediction rules in predicting the mortality of critically ill patients considered for intensive care unit (ICU) admission[48]. They concluded that ICU physicians more accurately discriminate between survivors and non-survivors in the first 24 hours of ICU admission than prediction rules do. It may thus be important to compare physicians' predictions with those of prediction rules, preferably already during the development phase of a prediction rule (requiring obviously a prospective design), but certainly in validation and impact studies. If physicians' predictions of probabilities have proven to outperform the probabilities provided by a rule, the rule will likely not be used in practice. In contrast, results in favour for the prediction rule can be used to convince physicians of using the rule when properly validated.

Second, prediction rules must have face validity; physicians must accept the logic, as well as the science of the rule. Clinical prediction rules that do not have face validity may not be applied in practice, even when effective[49].

Third, prediction rules may not be used because they are not user-friendly[3, 50]. The user-friendliness should be taken into account when developing the rule. A variable should only be considered as a potential predictor if obtaining this variable is also feasible in daily practice of the type of patients under study (not too time-consuming or costly). The user-friendliness of a prediction rule also depends on the way a prediction rule is presented; the original regression formula (as e.g. in the legend of Table 1) is the most exact and accurate form, but may involve cumbersome calculations requiring a calculator or computer. Although sumscores, risk stratification charts or nomograms may be less precise, they certainly are more user-friendly. With the introduction of electronic patient records, the use of regression formulas in daily practice will become much easier.

Finally, practical barriers may exist to act on the results of the prediction rule. For instance, when using a diagnostic prediction rule aiming to determine whether subsequent testing is necessary, such as the Ottawa ankle rules, physicians may be concerned about protecting themselves against litigation. Hence, they may still refer their patients for additional testing, while at the same time the prediction rule indicates that referral was not necessary[3]. Brehaut et al.[51] conducted a postal survey among 399 randomly selected physicians to examine whether physicians used the Ottawa ankle rules[37] for diagnosing ankle fractures. Most physicians (90%) reported to use the Ottawa ankle rules always or most of the time in appropriate circumstances, while only 42% actually based their decisions to order radiography primarily on the rule[51]. The same authors assessed why some prediction rules become widely used while others do not[52], taking the Canadian Cervical Spine Rule[53] as an example. They showed that older physicians and part-time working physicians were less likely to be familiar with the rule. The best predictors whether a rule would be used in practice were the familiarity acquired during training, the confidence in the usefulness of the rule, and the user-friendliness of the rule[52].

Final comments

We have given an overview of types of validation studies, of methods to improve or update a previously developed rule in case of disappointing accuracy in a validation study, and of aspects of impact studies and the implementation of prediction rules. A validated, and if necessary updated, rule may cautiously be applied in new patients that are similar to the patients in the development and validation populations. However, when the user has reasons to believe that the rule may perform differently in the new patients, data of the new patients should first be collected to test the accuracy, and preferably impact of the rule, before it rule is applied in daily clinical practice. Any rule may perform slightly different in a new patient sample due to sampling variation. In that situation, the rule does not need to be updated. The questions remains: when has a rule been sufficiently validated and updated? So far, this particular methodological area of prediction research has not been explored. Future research should address the question how many validation studies and what type of adjustments are needed before it is justified to implement a prediction rule into clinical practice.

Another subject of prediction research that may need more focus in the future is the methodology for systematic reviews and perhaps even meta-analyses of prediction rules for the same outcome[54, 55]. It will be a challenge to define how regression coefficients of prediction rules can be combined, and how to properly address publication bias; as prediction rules with good results are more likely to be published than rules with moderate results. Although the methodology for meta-analyses has been extensively described for etiologic and intervention studies, to our knowledge no research has been conducted for meta-analysis to combine several prediction rules.

Last, the potential gain in predictive accuracy and generalisability of a prediction rule developed on combined datasets with individual patient data from various studies on the same outcome (so-called individual patient studies) is a research area that needs more attention[56].

Our purpose was to stress the importance of testing the generalisability and impact of prediction rules, and outline the methods of such research. The relevance and importance of validating and testing the impact of prediction rules on physician's behaviour, and patients' outcome, has repeatedly been emphasized in the literature. Unfortunately, only a relatively small number of rules are validated, and hardly any study questions whether an implemented rule can change patient outcome. An increased focus on validation and impact studies will likely improve the application of valid prediction rules in daily clinical practice.

Acknowledgements

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZonMw 016.046.360; ZonMw 945-04-009).

References

1 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000 Feb 29;19(4):453-73.

2. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. Jama. 1997 Feb 12;277(6):488-94.

3. 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 Jul 5;284(1):79-84.
4. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006 Feb 7;144(3):201-9.
5. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985 Sep 26;313(13):793-9.
6. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001 Jul-Aug;8(4):391-7.
7. 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.
8. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999 Mar 16;130(6):515-24.
9. Copas JB. Regression, prediction and shrinkage. JR Stat Soc B. 1983;45:311-54.
10. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. American Statistician. 1983;37:36-48.
11. Harrell FE, Jr. Regression Modelling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.
12. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003 May;56(5):441-7.
13. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000 Apr 30;19(8):1059-79.
14. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001 Aug;54(8):774-81.
15. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003 Sep;56(9):826-32.
16. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999 Jul;16(1):9-13.
17. Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg. 1999 Jun;15(6):816-22; discussion 22-3.
18. Geissler HJ, Holzl P, Marohl S, Kuhn-Regnier F, Mehlhorn U, Sudkamp M, et al. Risk stratification in heart surgery: comparison of six score systems. Eur J Cardiothorac Surg. 2000 Apr;17(4):400-6.
19. Gogbashian A, Sedrakyan A, Treasure T. EuroSCORE: a systematic review of international performance. Eur J Cardiothorac Surg. 2004 May;25(5):695-700.
20. Kawachi Y, Nakashima A, Toshima Y, Arinaga K, Kawano H. Risk stratification analysis of operative mortality in heart and thoracic aorta surgery: comparison between Parsonnet and EuroSCORE additive model. Eur J Cardiothorac Surg. 2001 Nov;20(5):961-6.
21. Michel P, Roques F, Nashef SA. Logistic or additive EuroSCORE for high-risk patients? Eur J Cardiothorac Surg. 2003 May;23(5):684-7; discussion 7.
22. Nashef SA, Roques F, Hammill BG, Peterson ED, Michel P, Grover FL, et al. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American

cardiac surgery. Eur J Cardiothorac Surg. 2002 Jul;22(1):101-5.
23. Nilsson J, Algotsson L, Hoglund P, Luhrs C, Brandt J. Early mortality in coronary bypass surgery: the EuroSCORE versus The Society of Thoracic Surgeons risk algorithm. Ann Thorac Surg. 2004 Apr;77(4):1235-9; discussion 9-40.
24. Yap CH, Reid C, Yii M, Rowland MA, Mohajeri M, Skillington PD, et al. Validation of the EuroSCORE model in Australia. Eur J Cardiothorac Surg. 2006 Apr;29(4):441-6; discussion 6.
25. Nashef SA. Editorial comment EuroSCORE and the Japanese aorta. Eur J Cardiothorac Surg. 2006 Oct;30(4):582-3.
26. Oostenbrink R, Moons KG, Derksen-Lubsen G, Grobbee DE, Moll HA. Early prediction of neurological sequelae or death after bacterial meningitis. Acta Paediatr. 2002;91(4):391-8.
27. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol. 2002 May;20(2):96-107.
28. Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. J Clin Epidemiol. 2007 May;60(5):491-501.
29. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005 May;58(5):475-83.
30. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost. 2005 Jul;94(1):200-5.
31. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, et al. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. N Engl J Med. 2003 Sep 25;349(13):1227-35.
32. Di Nisio M, Squizzato A, Rutjes AW, Buller HR, Zwinderman AH, Bossuyt PM. Diagnostic accuracy of D-dimer test for exclusion of venous thromboembolism: a systematic review. J Thromb Haemost. 2007 Feb;5(2):296-304.
33. Bont J, Hak E, Hoes AW, Schipper M, Schellevis FG, Verheij TJ. A prediction rule for elderly primary-care patients with lower respiratory tract infections. Eur Respir J. 2007 May;29(5):969-75.
34. Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002 Dec;55(12):1201-6.
35. Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. Lancet. 1995 May 27;345(8961):1326-30.
36. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med. 2005 Jul 19;143(2):100-7.
37. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med. 1992 Apr;21(4):384-90.
38. Myers A, Canty K, Nelson T. Are the Ottawa ankle rules helpful in ruling out the need for x ray examination in children? Arch Dis Child. 2005 Dec;90(12):1309-11.
39. Altman DG. 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; 2007.
40. Hosmer DW, S L. Applied logistic regression. New York: John Wiley and Sons, Inc.; 1989.
41. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation. 1999 Apr 27;99(16):2098-104.
42. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation

and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004 Aug 30;23(16):2567-86.
43. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, vergouwe Y. Updating methods improved the predictive performance of clinical prediction models in new patients. J Clin Epidemiol. 2007;In Press.
44. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. Jama. 2001 Jul 11;286(2):180-7.
45. Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. Jama. 2004 Jun 2;291(21):2591-9.
46. Stiell I, Wells G, Laupacis A, Brison R, Verbeek R, Vandemheen K, et al. Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. Multicentre Ankle Rule Study Group. Bmj. 1995 Sep 2;311(7005):594-7.
47. Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. Community-Acquired Pneumonia Intervention Trial Assessing Levofloxacin. Jama. 2000 Feb 9;283(6):749-55.
48. Sinuff T, Adhikari NK, Cook DJ, Schunemann HJ, Griffith LE, Rocker G, et al. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. Crit Care Med. 2006 Mar;34(3):878-85.
49. Blackmore CC. Clinical prediction rules in trauma imaging: who, how, and why? Radiology. 2005 May;235(2):371-4.
50. Braitman LE, Davidoff F. Predicting clinical states in individual patients. Ann Intern Med. 1996 Sep 1;125(5):406-12.
51. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules „in the real world": how a widely disseminated rule is used in everyday practice. Acad Emerg Med. 2005 Oct;12(10):948-56.
52. Brehaut JC, Stiell IG, Graham ID. Will a new clinical decision rule be widely used? The case of the Canadian C-spine rule. Acad Emerg Med. 2006 Apr;13(4):413-20.
53. Stiell IG, Wells GA, Vandemheen KL, Clement CM, Lesiuk H, De Maio VJ, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. Jama. 2001 Oct 17;286(15):1841-8.
54. Altman DG. Systematic reviews of evaluations of prognostic variables. Bmj. 2001 Jul 28;323(7306):224-8.
55. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer. 2003 Apr 22;88(8):1191-8.
56. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. Eur J Obstet Gynecol Reprod Biol. 2003 Jun 10;108(2):121-5.

Chapter 3.2

Validation of a clinical prediction rule for severe postoperative pain in new settings

Abstract

Background: Recently, a prediction rule has been derived to preoperatively predict the risk of severe pain in the first postoperative hour in surgical inpatients. We aimed to modify the rule to enhance its use in both inpatients and outpatients. Subsequently, we prospectively tested the modified rule in patients that underwent surgery later in time and in another hospital (external validation).

Methods: The rule was developed on the data of 1395 adult inpatients and modified with the data of 549 outpatients that underwent surgery between 1997 and 1999 in the Academic Medical Center Amsterdam, the Netherlands. We tested the performance of the rule in 1035 patients that underwent surgery in 2004 in the University Medical Center Utrecht, the Netherlands. Modification of the prediction rule included reclassification of the predictor 'type of surgery' and addition of interaction terms between surgical setting (ambulatory surgery: yes/no) and the other predictors. We studied the performance by assessing the calibration (agreement between observed frequencies and predicted risks) and the discrimination (ability to distinguish between patients at high and low risk).

Results: One third of the patients in the Utrecht cohort reported severe postoperative pain (36%), compared to 62% of the patients in the Amsterdam cohort. The distribution of most predictors was similar in the two cohorts, although the patients in the Utrecht cohort were slightly older, underwent more often ambulatory surgery and had less often large expected incision sizes than patients in the Amsterdam cohort. The prediction rule showed good calibration and reasonable discrimination (ROC area 0.65 (95% confidence interval 0.57 - 0.73)).

Conclusions: A previously developed prediction rule to predict severe postoperative pain was modified to enhance its valid use in both inpatients and outpatients. By validating the rule in patients that underwent surgery in another hospital and more recently, the rule proofed to be generalisable in place and time. If this prediction rule proves to be robust in other hospitals, it may improve the quality of acute postoperative pain management by timely identifying patients who will benefit from preventive treatments.

Introduction

Moderate to severe acute postoperative pain occurs frequently after a variety of surgical procedures. Incidences of up to 50% in inpatients and 40% in outpatients (patients undergoing ambulatory surgery) have been reported.[1-4] Severe postoperative pain may result in patient discomfort, reduced patient satisfaction, delayed discharge from the postoperative anesthesia care unit and hospital, and limited mobility and return to normal activities.[5] Moreover, it can promote delirium in the elderly[6] and may develop into chronic pain syndromes.[7]

Prediction rules for various postoperative outcomes such as mortality have been developed and are being used for risk management. Surprisingly, no rules existed for preoperative estimation of the risk of acute or late postoperative pain. Such rules could preoperatively determine patients at high risk versus low risk to help direct appropriate preventive pain treatment. Given the reported high incidences of severe acute postoperative pain, the approach to prevent pain in current practice apparently seems insufficient to timely identify and treat patients at high risk.[2,8]

Recently, a multivariable prediction rule (including only predictors that are easy to obtain during a preoperative visit) was derived to preoperatively predict the risk of severe pain in the first postoperative hour in surgical inpatients.[9] It would be useful if this rule could predict severe acute postoperative pain also in outpatients. We therefore modified the inpatients rule to be used in inpatients and outpatients, for settings with different incidences (prior probabilities) of postoperative pain. Subsequently, we prospectively tested this modified rule in patients that underwent surgery later in time in another hospital (external validation). We used state of the art methods for modification and validation of clinical prediction rules that can be used for prediction rules for any clinical problem. Hence, this paper may also serve as a methodological illustration on the modification and validation of clinical prediction rules in general.

Patients and Methods

The prediction rule for inpatients

The prediction rule for severe acute postoperative pain was originally derived with data from a cohort of 1416 surgical inpatients that underwent surgery in the Academic Medical Center Amsterdam, the Netherlands (further referred to as the Amsterdam cohort).[9] In brief, the patients that underwent surgery were aged 18-85. All types of surgery were included, except cardiac surgery and intracranial neurosurgical procedures. Exclusion criteria were emergency surgery, pregnancy, ASA physical status 4 and morbid obesity (weight > 120 kg). Induction of anesthesia was achieved with thiopental in patients randomized to isoflurane/nitrous oxide, and with propofol in patients randomized to total intravenous anesthesia with propofol/air. Anesthesia was maintained with propofol or isoflurane in nitrous oxide according to the randomization. Intra-operatively, the anesthesiologist was free to use opioid analgesics and muscle relaxants as needed.[10] Patients did not receive preventive analgesics and only in a few patients intra-operative opioids were administered.

The outcome of the prediction rule was the presence of severe acute postoperative pain. It was defined as a numerical rating scale (NRS) score equal to or higher than 8 (where 0 indicates no pain at all, and 10 the most severe pain imaginable), occurring at least once within the first hour at the postoperative anesthesia care unit. A trained

and blinded research nurse recorded the severity of pain every 15 minutes with a NRS. If patients were not awake they received a NRS score of 0 (no pain) for that time point.

The prediction rule was presented as a formula and as an easy to use nomogram (see Appendix). The included predictors were gender, age, type of surgery (ophthalmology, laparoscopy, ear/nose/throat, orthopedic surgery, intra-abdominal and other type of surgery), expected incision size ≥ 10 cm, a preoperative pain score, an anxiety score and a need for information score. The predictors anxiety and need for information score were based on the APAIS questionnaire, which consists of six questions, each scored on a 5-point Likert scale from 1 (not at all) to 5 (extremely). The APAIS is specifically designed to assess the patients' preoperative anxiety score (4 questions, score range 4-20) and a information seeking behavior score to assess the patients' need for information regarding the scheduled surgery and anesthesia (2 questions, score range 2-10).[11]

Modifying the prediction rule for surgical outpatients

In the present analysis, we used the data of the 549 surgical outpatients that participated in the Amsterdam cohort of the original study from which the inpatients emerged. The same methods of data collection were applied. Hence, the same predictors and outcome definitions could be used. As the prediction rule performed insufficiently in these outpatients, we modified the rule so it was applicable and valid to both inpatients and outpatients. To improve the performance, we made three adjustments to the rule.

First, we used a more widely accepted definition of severe acute postoperative pain for both in- and outpatients, i.e. NRS score ≥ 6, to better adhere to the indication for administering acute pain treatment in current practice.[12]

Second, the classification of type of surgery that was used in the rule for inpatients was initially developed for prediction of post operative nausea and vomiting rather than for acute postoperative pain.[9] Since for the latter purpose no suitable classification could be found in the literature, we developed this classification. We identified twenty-seven groups of surgical procedures, based on clinical experience, current practice and interviews with surgeons and anesthesiologists (Table 1). Subsequently, the univariable association between each surgical group and severe acute postoperative pain was estimated. Groups with similar associations were further combined.

Third, as we expected that inpatients may have a higher risk of postoperative pain than outpatients, we included an additional predictor 'type of patient' indicating surgical setting (ambulatory surgery: yes versus no).

Subsequently, the regression coefficients of the seven predictors of the original rule (see Appendix) plus surgical setting were estimated in a multivariable logistic regression model. We a priori hypothesized that the predictive effect of some of these seven predictors could be different for surgical inpatients and outpatients. We tested this with interaction terms between the seven predictors and type of patient.

As we adjusted the threshold of severe postoperative pain from a NRS ≥ 8 to a NRS ≥ 6, other variables than included in the original rule could have become important. Hence, we performed an extra analysis to quantify whether other variables – such as body mass index, duration of surgery and type of anesthesia (intravenous versus inhalation) – had an added predictive effect. All these other variables were far from significant and no important predictors were missed.

Table 1 Surgical procedures conducted in patients of the Amsterdam cohort, ordered by increasing incidence of severe acute postoperative pain (defined as ≥ 6 on a numerical rating scale).

Surgical procedure	Incidence %	Severe pain N	(total N)
Lowest expected pain			
Endoscopic urology	26	7	(27)
Testical surgery (including orchidopexy, biopsy, prosthesis implantation, vasoepididymostomy, testis-scrotum exploration)	27	3	(11)
Eye surgery (including strabismus)	37	43	(116)
Low expected pain			
Pharyngo- and laryngoscopy plus biopsy	40	8	(20)
Ear Nose Throat surgery	47	130	(277)
Diagnostic laparoscopy	48	50	(105)
Gynaecologic surgery (non-abdominal non-laparoscopic)	49	34	(69)
Minor rectal surgery	49	18	(37)
Oral soft tissue surgery	55	21	(38)
Carotid endarterectomy	56	5	(9)
Moderate expected pain			
Skin surgery or lymph node biopsy	58	43	(74)
Peripheral vascular procedures (including varicose veins)	59	26	(44)
Minor breast surgery	61	39	(64)
Procedures on muscle and/or ligaments of extremities	63	75	(119)
Upper abdominal surgery with epidural, including hepato-billiary, oesophageal, pancreatic and intestinal surgery	63	19	(30)
High expected pain			
Major breast surgery	67	45	(67)
Bone procedures, including cranial/facial, oral, spine, orthopedic/ traumatology procedures on clavicle, extremities, hip and pelvis. Instrumentation or removal of instrumentation, including spine, hip, jaw/denture, hand/wrist, clavicle, elbow, ankle/foot or knee. Arthroscopy of shoulder, hip/pelvis and extremities.	68	255	(377)
Procedures for abdominal wall herniation	69	42	(61)
Nefrectomy	69	9	(13)
Highest expected pain			
Therapeutic laparoscopic procedures, including laparoscopic cholecystectomy, gynaecologic laparoscopy and other therapeutically laparoscopy	76	94	(123)
Intra abdominal surgery without epidural, including colon, bladder, prostate, vascular and gynaecological surgery	80	49	(61)
Tonsillectomy (in patients over 16 year)	80	37	(46)
Herniated disc surgery	84	16	(19)
Thyroid procedures	86	12	(14)
Peripheral nerve reconstruction	92	12	(13)
Vaginal hysterectomy	100	7	(7)

The incidence of severe postoperative pain might be different in other patient populations. We therefore estimated what the effect of the different incidences would be on the intercept, and presented these adjusted intercepts with the prediction rule.

Prediction rules usually show too optimistic performance in the patients from which they are derived.[13-16] We therefore estimated the amount of overoptimism and adjusted the modified rule for it, using bootstrapping techniques.[14,17,18]

External validation

The predictive performance of prediction rules always needs to be tested in new patients before they can be applied in daily clinical practice.[13-16] The inpatients and outpatients in the Amsterdam cohort underwent surgery between April 1997 and January 1999. To test whether the rule was generalisable across time and place, we studied the predictive performance of the rule in 1035 new consecutive patients (validation set) from a prospective cohort that underwent surgery between February and December 2004 in a different academic hospital (University Medical Center Utrecht, the Netherlands, further referred to as the Utrecht cohort). The same predictors and outcome definitions and measurements were used as in the original Amsterdam cohort.

Predictive performance measures

To study the predictive performance of the modified prediction rule, we assessed the calibration and discrimination of the rule in the Utrecht cohort. Calibration refers to the agreement between the predicted risks and observed frequencies of severe acute postoperative pain in the new patients. This was graphically assessed with a calibration plot.[14] Discrimination is the ability of the rule to distinguish between patients with severe pain and patients without severe pain, and was quantified with the area under the ROC curve (ROC area). An ROC area ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination).[14]

Results

The modified prediction rule for surgical inpatients and outpatients

The new classification of type of surgery resulted in five categories (Table 1); i.e. lowest expected incidence of pain (observed incidence of severe postoperative pain 33%), low expected incidence of pain (47%), moderate expected incidence of pain (61%), high expected incidence of pain (68%) and highest expected incidence of pain (84%). Twenty-one patients with an unknown or rare surgical procedure (less than 5 patients, i.e. surgery of the penis, bone marrow, and trachea) were excluded from the analysis, since their effect on postoperative pain could not be reliably estimated. Some surgical groups still had small numbers (more than 5 but less then 20 patients, such as carotid endarterectomy and vaginal hysterectomy, Table 1). These groups were explicitly retained in the analysis to enhance generalisibility of the final results, even though the uncertainty in the outcome incidence will be higher for these groups than in groups with, for example, more than 100 patients. The effect of gender and type of surgery was different for inpatients and outpatients and these interaction terms were included in the prediction rules.

Of the surgical outpatients, 48% (265/549) reported a NRS score ≥ 6 within the first hour after arriving at the postoperative anesthesia care unit, versus 67% (935/1395) of the inpatients (Table 2, second and third column). Outpatients were on average younger and had less often a large incision (20 versus 44%). Outpatients less often underwent types of surgery with high or highest expected pain than inpatients.

The model was presented as a formula (Table 3) and as an easy to use score chart (Figure 1). Table 3 shows the intercept and the regression coefficients of the predictors in the modified prediction rule. The risk of severe postoperative pain was increased by a large expected incision size (larger than 10 cm), high preoperative pain and anxiety scores, and decreased by female gender, higher age and a higher need for information

Table 2 Distribution of the patient characteristics of the patients that underwent surgery between April 1997 and January 1999 in the Amsterdam cohort, and the patients that underwent surgery between February and December 2004 in the Utrecht cohort; % (n) unless stated otherwise.

Patient characteristics	Inpatients (n=1395)		Outpatients (n=549)		Inpatients (n=591)		Outpatients (n=444)	
Female gender	58	(810)	55	(300)	53	(312)	64	(284)
Age (years)*	45	(15)	38	(12)	50	(17)	42	(14)
Preoperative pain score*	3.1	(2.9)	2.7	(2.6)	3.4	(3.0)	3.3	(2.8)
Type of surgery								
Lowest expected incidence of pain	4	(51)	4	(21)	9	(54)	9	(38)
Low expected incidence of pain	27	(379)	38	(211)	28	(165)	30	(132)
Medium expected incidence of pain	15	(214)	24	(134)	18	(104)	22	(96)
High expected incidence of pain	36	(504)	30	(167)	30	(180)	36	(159)
Highest expected incidence of pain	18	(247)	3	(16)	15	(88)	4	(19)
Expected incision size ≥ 10 cm	44	(614)	20	(110)	11	(65)	0	(0)
APAIS anxiety score*	9.4	(4.0)	9.6	(4.1)	9.2	(3.4)	9.2	(3.5)
APAIS need for information*	6.6	(2.2)	6.4	(2.2)	5.4	(1.7)	6.2	(2.3)
Severe acute postoperative pain, NRS ≥ 6	67	(935)	48	(265)	36	(215)	33	(147)

* Mean (standard deviation)

scores. The interaction terms indicated that the effect of gender was different in inpatients and outpatients. Also, outpatients scheduled for a same type of surgery as inpatients had a lower risk of severe postoperative pain (indicated by the negative interaction term) than the inpatients.

We estimated the intercepts for other incidences of severe postoperative pain (Table 3), so that when the rule is applied in a population with other incidences of severe postoperative pain, the intercept of the rule can be adjusted. When the incidence of severe postoperative pain is lower, the intercept needs to be more negative, leading to lower predicted risks.

The calibration of the modified rule was good (plot not shown). The ROC area of the rule was 0.71 (95% confidence interval: 0.66,0.76).

External validation of the modified rule in the Utrecht cohort

One third of the patients in the Utrecht cohort reported severe postoperative pain (36%), compared to 62% of the patients in the Amsterdam cohort (Table 2). The distribution of most predictors was similar in the two cohorts, although the patients in the Utrecht cohort were slightly older, underwent more often ambulatory surgery and had less often large expected incision sizes than patients in the Amsterdam cohort. Inpatients had a similar incidence of severe postoperative pain as outpatients in the Utrecht cohort (36% and 33%). We tested the modified prediction rule with an intercept for an incidence of 35% (-1.53 in stead of -0.42), corresponding to the lower incidence of severe postoperative pain in the Utrecht cohort (Table 3, Figure 1).

Table 3 Intercept and regression coefficients of the predictors in the modified prediction rule estimated in the Amsterdam cohort (n = 1944). Different incidence specific intercepts are shown to be used in populations with different incidences of severe postoperative pain than the incidence in the Amsterdam cohort.

Predictor	Regression coefficient
Female gender	-0.004
Age	-0.009
Type of surgery	
Lowest expected pain (reference)	-
Low expected pain	0.50
Moderate expected pain	0.92
High expected pain	1.05
Highest expected pain	1.72
Expected incision size ≥ 10 cm	0.39
Preoperative pain score	0.11
Anxiety score	0.05
Need for information score	-0.05
Ambulatory surgery	-0.70
Ambulatory surgery * female gender	0.67
Ambulatory surgery * type of surgery	
Ambulatory surgery * lowest expected pain (reference)	-
Ambulatory surgery * low expected pain	-0.10
Ambulatory surgery * moderate expected pain	-0.47
Ambulatory surgery * high expected pain	-0.07
Ambulatory surgery * highest expected pain	-1.51
Intercept	-0.42
Intercept adjusted for prior probability	
60%	-0.33
55%	-0.53
50%	-0.73
45%	-0.93
40%	-1.14
35%	-1.35
30%	-1.58
25%	-1.83
20%	-2.12

The formula of the prediction model is of the form:

$$\log\left(\frac{\text{risk of pain}}{1-\text{risk of pain}}\right) = \text{linear predictor} = ß_0 + ß_1*\text{predictor}_1 + ß_2*\text{predictor}_2 + \ldots + ß_n*\text{predictor}_n$$

In this formula, $ß_0$ is the intercept and $ß_1$ till $ß_n$ are the regression coefficients. The risk of severe postoperative pain in individual patients (scale: 0-100%) can be calculated with the formula:

$$\text{risk} = \frac{1}{1+e^{-\text{linear predictor}}}$$

Figure 1 Score chart to predict the risk of severe acute postoperative pain for inpatients and outpatients. The scores are based on the regression coefficients of the prediction rule. For each patient, a sumscore can be calculated by counting the scores that correlate to the characteristics of the patient. The total sumscore can be linked to the individual risk using the box. For patients in settings with a different incidence (prior probability) of severe postoperative pain, the corresponding constant should be used. For example, in an outpatient setting (score = -4), a female patient (score = 3) of age 43 (score = -2), has a preoperative pain score of 9 (score = 5) is scheduled for a high expected risk of pain procedure (score = 5) with a small expected incision size (score = 0), and has a preoperative anxiety score of 16 (score = 4) and a preoperative need for information score of 4 (score = -1). This patient has a total sum score of 10. This total sum score refers to a risk of severe postoperative pain of 83% (using the lower part of the figure). When we use the formula of the prediction rule to estimate the risk of severe postoperative pain for this patient (Table 3), we find a risk of 85%.

Predictor	**Score per predictor**		
	Inpatient	**Outpatient**	
Female gender	0	3	…
Age			…
< 35	-1		
35-57	-2		
58-79	-3		
≥ 80	-4		
Pre operative pain score			…
1 or 2	1		
3 or 4	2		
5 or 6	3		
7 or 8	4		
9	5		
10	6		
Large incision size (≥ 10 cm)	2		…
Type of surgery			…
Lowest pain	0	0	
Low pain	3	2	
Moderate pain	5	2	
High pain	5	5	
Highest pain	9	1	
Pre operative anxiety			…
4, 5 or 6	1		
7, 8, 9 or 10	2		
11, 12, 13 or 14	3		
15, 16, 17 or 18	4		
19 or 20	5		
Pre operative need for information			…
2, 3, 4 or 5	-1		
6, 7, 8 or 9	-2		
10	-3		
Constant	0	-4	…
Constant for other incidences of severe postoperative pain			
60%	0	-4	
55%	-2	-6	
50%	-3	-7	
45%	-4	-8	
40%	-5	-9	
35%	-6	-10	
30%	-7	-11	
25%	-8	-12	
20%	-9	-15	
Total sum score			…

Total sumscore	< -8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12
Risk (%)	19	22	25	29	30	31	33	37	42	49	52	57	61	67	70	74	77	81	83	84	87

Figure 2 shows the calibration line of the modified prediction rule in the Utrecht cohort. The dotted line shows the ideal situation in which the predicted risks and the observed frequencies of postoperative pain are completely in agreement. The solid line shows the observed association between the predicted risks and the observed frequencies.

The prediction rule showed good calibration when the rule was tested in the patients of the Utrecht cohort (Figure 2a). Note that when the predicted risks were higher than 60%, the predicted risks were slightly too high. To illustrate the necessity to adjust the intercept according to the lower incidence of severe postoperative pain, we also presented the calibration line when we would have used the original intercept of -0.42 (Figure 2b). The predicted risks would be systematically higher than the observed frequencies.

Figure 2 Calibration line of the modified prediction rule in the Utrecht cohort with adjusted intercept corresponding to the different incidence (prior probability) of severe postoperative pain (35% versus 62% in the Amsterdam cohort) (a) and with the original intercept that corresponds to the incidence in the Amsterdam cohort (b). Triangles indicate the observed frequency of severe acute postoperative pain per decile of predicted risk. The solid line shows the relation between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration, in which the predicted risks equal the observed frequencies of severe postoperative pain.

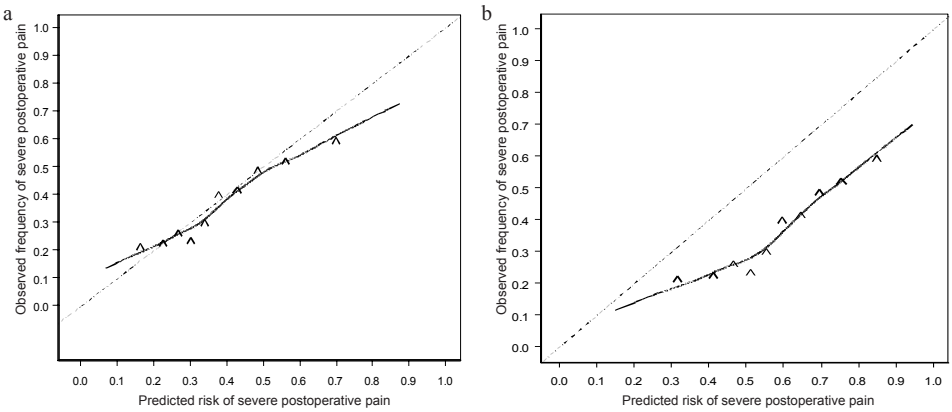


Table 4 shows the observed number of patients with and without severe postoperative pain across score and risk categories estimated by the modified rule in the Utrecht cohort. The incidence of severe postoperative pain per risk stratum increased from 18% to 65%. For example, for patients with a score of -10, the observed incidence of severe postoperative pain was 21%, while for patients with a score of 1, the observed incidence was 65%. Again, when the predicted risks were higher than 60%, the predicted risks were slightly too high (see also Figure 2a).

The ROC area of the modified prediction rule in the Utrecht cohort was 0.65 (0.57 - 0.73).

Table 4 Calculated risk and observed incidence of severe postoperative pain for different score thresholds in the Utrecht cohort. The score is calculated with the score chart of the modified prediction rule.

Calculated score*	< -15	-14 to -10	-9 to -7	-6 to -5	-4 to 0	> 0
Mean estimated risk† of severe pain, %	19	26	32	40	54	75
Observed risk of severe pain, %(n)	18 (9)	21 (54)	34 (84)	42 (84)	48 (112)	65 (33)
Number of patients	49	252	250	200	233	51

* Categories of the score as calculated from the score chart (Figure 1)
† Risk of severe acute postoperative pain as estimated by the modified prediction rule

Discussion

We modified a previously developed rule to preoperatively predict the risk of severe acute postoperative pain in surgical inpatients, to make it applicable to both inpatients and outpatients. We used a more widely accepted definition of severe acute postoperative pain, i.e. NRS $\geq$ 6, to better adhere to the indication for administering pain treatment in current practice. We revised the surgical classification and added surgical setting. We validated the modified prediction rule in inpatients and outpatients from another cohort that underwent surgery more recently and in another hospital. Since the incidence of severe postoperative pain was lower in this cohort (36% versus 62%), we used an intercept that corresponds to an incidence of 35%. The modified rule showed good calibration and reasonable discrimination (ROC area = 0.65).

The data to modify the prediction rule were obtained from patients that underwent surgery between 1997 and 1999. Since then, surgical protocols may have changed. For example, types of surgery that used to be performed on inpatients only may nowadays be applied in ambulatory surgery as well. Also, the incidence of severe postoperative pain may have changed over time due to changing treatment protocols. Therefore, we specifically tested the modified rule in patients from another hospital, that underwent surgery more recently and that had a lower incidence of severe postoperative pain. The modified rule showed good predictive performance. Although interest in the field may be shifting to include consideration of late postoperative pain, we focused on acute postoperative pain for two reasons. Firstly, the occurrence of acute postoperative pain remains an important outcome for patients. Secondly, it has become a quality of care indicator in various countries.

Type of surgery was an important predictor in the original prediction rule for inpatients. However, the surgical classification that was used was developed for the prediction of postoperative nausea and vomiting. It was most likely not sensitive enough for the prediction of postoperative pain. Previous presented classifications in the literature of type of surgery for the prediction of postoperative pain were not suitable for our study, as they did not cover all surgical procedures that were conducted in our populations.[19] Therefore, we developed a new classification to address this issue. However, validation of the surgical classification is required, especially since some groups of surgical procedures were relatively underrepresented in the surgical classification.

We found different effects of gender and type of surgery for inpatients versus outpatients. This was illustrated by the regression coefficients for interaction effect (unequal to 0) (Table 3). Gender has no predictive effect in inpatients, whereas in outpatients the risk

was twice as high for female patients than male patients. For all types of surgery, the risk of severe acute postoperative pain is lower for outpatients than for inpatients, as shown by the negative regression coefficients for the interaction terms. This indicates that an outpatient scheduled for the same type of surgery has a lower risk than an inpatient (with the same characteristics). This was also shown when the entire analysis was repeated for inpatients and outpatients separately. A simple explanation can not be given for this phenomenon. Probably, an inpatient and outpatient scheduled for the same type of surgery do differ on other characteristics, which were not fully covered by the predictors in our rule. Probably the severity of the disorder requiring surgery may still be different between in- and outpatients.

The effects of most predictors included in the modified rule have been described before, such as the risk increasing effect of high preoperative pain scores,[20] the risk increasing effect of preoperative anxiety[21-23] and the risk decreasing effect of age.[19,20,24] The effect of gender remains subject of debate. Our result confirmed the higher effect of female gender on postoperative pain in outpatients.[12] Although there are studies that found an enhancing effect of female gender in inpatients,[20,25] most studies (including ours) found no or just a small effect of female gender in inpatients.[24,26,27] Type of surgery has always been known as an important predictor of postoperative pain.[20,28-31] We found that expected incision size is an independent predictor additional to type of surgery. It may seem peculiar to include a predictor of which the value can only be observed postoperatively in a rule that is applied preoperatively. However, in non-emergency surgery, expected incision size can be reliably estimated beforehand, given the large number of detailed surgical protocols in current practice. We are not aware of other studies that examined the predictive effect of expected incision size on the risk of postoperative pain.

A few other potential predictors of severe postoperative pain that were not included in our rule are described in the literature, notably body mass index (BMI)[24,28] and duration of surgery,[22,24,26,28,29] although with conflicting results. In the original study among the surgical inpatients, duration of surgery and BMI had no independent predictive value. Also in an additional analysis among in- and outpatients combined, these factors – including type of anesthesia – showed no additional predictive effect.

Preoperative pain tests, such as cold and thermal stimuli, suprathreshold pain stimuli and burn tests, have shown to predict the occurrence of acute postoperative pain.[24,32-36] However, such tests have less applicability in routine clinical care as they may be time consuming for the doctor and burdensome for patients. Yet, their added predictive value beyond the more easily obtainable predictors included in our rule remains to be quantified.

Preventive analgesic treatment should be considered for patients at high risk of severe acute postoperative pain. For the proper use of the modified prediction rule in clinical practice, a specific risk threshold for the indication of preventive treatment has to be chosen. Although practitioners can infinitive choose their desired threshold, we note that when a high risk threshold is chosen, fewer patients will receive preventative pain treatment (saving unnecessary side effects and treatment costs). This is at the expense that some patients that needed treatment, are not treated (which may in turn create additional costs). In contrast, when a low risk threshold is chosen, more patients will receive preventative pain treatment, which will reduce the incidence of severe acute postoperative pain, yet at the expense of over treatment (potentially leading to unnecessary side effects and higher costs).

With many prediction rules available for surgical complications such as myocardial infarction and mortality, surprisingly no feasible rules exist for the risk of postoperative pain. One prediction rule for postoperative pain was derived on patients undergoing orthopedic and intraperitoneal surgery.[22] This rule included anesthetic technique, expectation of postoperative pain and chronic sleeping difficulties. However, only chronic sleeping difficulties showed a predictive effect in the validation set. Moreover, the rule was developed on a small dataset (304 patients, with 153 experiencing severe postoperative pain) and far too many predictors (n=62) were examined. In general, at least ten events are needed for each considered predictor.[14] This means that at least 620 patients with severe postoperative pain were needed instead of 153. Therefore, one may question the predictive value of this rule when applied in new patients.

In conclusion, a previously developed prediction rule to predict severe postoperative pain was modified and externally validated in inpatients and outpatients. By presenting intercepts for different incidences (prior probabilities) of severe postoperative pain, the rule could be applied to a patient population with a different incidence. The rule proofed to be generalisable in place and time, when tested in patients that underwent surgery in another hospital and more recently. If this prediction rule proves to be robust in other hospitals, it could improve the quality of acute postoperative pain management by timely identifying patients who will benefit from preventive treatments.

Acknowledgement

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZonMw 016.046.360).

References

1 Chauvin, M. State of the art of pain treatment following ambulatory surgery. Eur.J.Anaesthesiol. 20, 3-6. 2003.

2 Huang, N., Cunningham, F., Laurito, C. E., and Chen, C. Can we do better with postoperative pain management? Am.J.Surg. 182[5], 440-448. 2001.

3 Shaikh, S., Chung, F., Imarengiaye, C., Yung, D., and Bernstein, M. Pain, nausea, vomiting and ocular complications delay discharge following ambulatory microdiscectomy. Can. J.Anaesth. 50[5], 514-518. 2003.

4 Svensson, I., Sjostrom, B., and Haljamae, H. Assessment of pain experiences after elective surgery. J.Pain Symptom.Manage. 20[3], 193-201. 2000.

5 Myles, P. S., Williams, D. L., Hendrata, M., Anderson, H., and Weeks, A. M. Patient satisfaction after anaesthesia and surgery: results of a prospective survey of 10,811 patients. Br.J.Anaesth. 84[1], 6-10. 2000.

6 Lynch, E. P., Lazor, M. A., Gellis, J. E., Orav, J., Goldman, L., and Marcantonio, E. R. The impact of postoperative pain on the development of postoperative delirium. Anesth.Analg. 86[4], 781-785. 1998.

7 Katz, J., Jackson, M., Kavanagh, B. P., and Sandler, A. N. Acute pain after thoracic surgery predicts long-term post-thoracotomy pain. Clin.J.Pain 12[1], 50-55. 1996.

8 Apfelbaum, J. L., Chen, C., Mehta, S. S., and Gan, T. J. Postoperative pain experience: results from a national survey suggest postoperative pain continues to be undermanaged. Anesth. Analg. 97[2], 534-40. 2003.

9 Kalkman, C. J., Visser, K., Moen, J., Bonsel, G. J., Grobbee, D. E., and Moons, K. G. Preoperative

prediction of severe postoperative pain. Pain 105[3], 415-423. 2003.
10 Visser, K., Hassink, E. A., Bonsel, G. J., Moen, J., and Kalkman, C. J. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea with vomiting and economic analysis. Anesthesiology 95[3], 616-626. 2001.
11 Moerman, N., van Dam, F. S., Muller, M. J., and Oosting, H. The Amsterdam Preoperative Anxiety and Information Scale (APAIS). Anesth.Analg. 82[3], 445-451. 1996.
12 Rosseland, L. A. and Stubhaug, A. Gender is a confounding factor in pain trials: women report more pain than men after arthroscopic surgery. Pain 112[3], 248-253. 2004.
13 Altman, D. G. and Royston, P. What do we mean by validating a prognostic model? Stat.Med. 19[4], 453-473. 29-2-2000.
14 Harrell, F. E., Jr., Lee, K. L., and Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat.Med. 15[4], 361-387. 28-2-1996.
15 Justice, A. C., Covinsky, K. E., and Berlin, J. A. Assessing the generalizability of prognostic information. Ann.Intern.Med. 130[6], 515-524. 16-3-1999.
16 Reilly, B. M. and Evans, A. T. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann.Intern.Med. 144[3], 201-209. 7-2-2006.
17 Efron B. Censored Data and the Bootstrap. Journal of the American Stat Association 76[374], 312-319. 1981.
18 Steyerberg, E. W., Harrell, F. E., Jr., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., and Habbema, J. D. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J.Clin.Epidemiol. 54[8], 774-781. 2001.
19 Macintyre, P. E. and Jarvis, D. A. Age is the best predictor of postoperative morphine requirements. Pain 64[2], 357-364. 1996.
20 Thomas, T., Robinson, C., Champion, D., McKell, M., and Pell, M. Prediction and assessment of the severity of post-operative pain and of satisfaction with management. Pain 75[2-3], 177-185. 1998.
21 Chapman, C. R. Psychological aspects of pain patient treatment. Arch.Surg. 112[6], 767-772. 1977.
22 Mamie, C., Bernstein, M., Morabia, A., Klopfenstein, C. E., Sloutskis, D., and Forster, A. Are there reliable predictors of postoperative pain? Acta Anaesthesiol.Scand. 48[2], 234-242. 2004.
23 Scott, L. E., Clum, G. A., and Peoples, J. B. Preoperative predictors of postoperative pain. Pain 15[3], 283-293. 1983.
24 Bisgaard, T., Klarskov, B., Rosenberg, J., and Kehlet, H. Characteristics and prediction of early pain after laparoscopic cholecystectomy. Pain 90[3], 261-269. 15-2-2001.
25 Aubrun, F., Salvi, N., Coriat, P., and Riou, B. Sex- and age-related differences in morphine requirements for postoperative pain relief. Anesthesiology 103[1], 156-160. 2005.
26 Kotzer, A. M. Factors predicting postoperative pain in children and adolescents following spine fusion. Issues Compr.Pediatr.Nurs. 23[2], 83-102. 2000.
27 Logan, D. E. and Rose, J. B. Gender differences in post-operative pain and patient controlled analgesia use among adolescent surgical patients. Pain 109[3], 481-487. 2004.
28 Chung, F., Ritchie, E., and Su, J. Postoperative pain in ambulatory surgery. Anesth.Analg. 85[4], 808-816. 1997.
29 Dahmani, S., Dupont, H., Mantz, J., Desmonts, J. M., and Keita, H. Predictive factors of early morphine requirements in the post-anaesthesia care unit (PACU). Br.J.Anaesth. 87[3], 385-389. 2001.

30 Ellstrom, M., Olsen, M. F., Olsson, J. H., Nordberg, G., Bengtsson, A., and Hahlin, M. Pain and pulmonary function following laparoscopic and abdominal hysterectomy: a randomized study. Acta Obstet.Gynecol.Scand. 77[9], 923-928. 1998.

31 Nguyen T, Malley R, Inkelis SH, and Kuppermann N. Comparison of prediction models for adverse outcome in pediatric meningococcal disease using artificial neural network and logistic regression analyses. 2002. Journal of Clinical Epidemiology.

32 Granot, M., Lowenstein, L., Yarnitsky, D., Tamir, A., and Zimmer, E. Z. Postcesarean section pain prediction by preoperative experimental pain assessment. Anesthesiology 98[6], 1422-1426. 2003.

33 Hsu, Y. W., Somma, J., Hung, Y. C., Tsai, P. S., Yang, C. H., and Chen, C. C. Predicting postoperative pain by preoperative pressure pain assessment. Anesthesiology 103[3], 613-618. 2005.

34 Kehlet, H., Rung, G. W., and Callesen, T. Postoperative opioid analgesia: time for a reconsideration? J.Clin.Anesth. 8[6], 441-445. 1996.

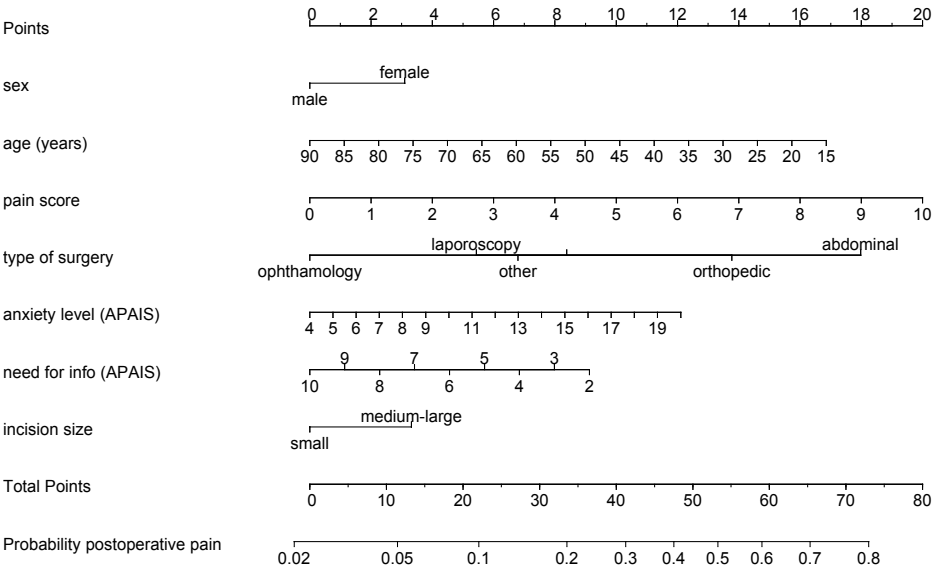
35 Kim, Hyungsuk, Neubert, John K., Rowan, Janet S., Brahim, Jaime S., Iadarola, Michael J., and Dionne, Raymond A. Comparison of experimental and acute clinical pain responses in humans as pain phenotypes. The Journal of Pain 5[7], 377-384. 2004.

36 Werner, M. U., Duun, P., and Kehlet, H. Prediction of postoperative pain by preoperative nociceptive responses to heat stimulation. Anesthesiology 100[1], 115-119. 2004.

Appendix

Figure 1 Nomogram and formula of the original prediction rule to predict the probability of severe postoperative pain within the first hour after surgery in surgical inpatients.

1a. Nomogram



1b. Formula

$$\log\left(\frac{\text{risk of pain}}{1-\text{risk of pain}}\right) = \text{linear predictor} = -1.74 + 0.22*\text{female gender} - 0.016*\text{age} + 0.38*\text{laparoscopy} + 0.59*\text{ear/nose/throat surgery} + 0.97*\text{orthopedic surgery} + 1.37*\text{intra-abdominal surgery} + 0.48*\text{other type of surgery} + 0.23*\text{expected incision size} \geq 10\text{ cm} + 0.14*\text{preoperative pain score (NRS)} + 0.053*\text{APAIS anxiety score} - 0.080*\text{APAIS need for information score}.$$

Chapter 3.3

A simple method to update clinical prediction rules

Abstract
Clinical prediction rules need to be tested in new patients (validated) before they can be applied in daily clinical practice. However, many prediction rules are developed without being validated in new patients. And when they are, investigators often develop a new rule when the predictive performance seems insufficient in the new patients. As a result, the information that is captured in the original rule is wasted. A better strategy is to simply update the rule, by combining the information of the original rule with the information of the new patients.
We present an example of a prediction rule that was validated and subsequently updated to adjust it to the new patients. The prediction rule was developed to predict the risk of severe postoperative pain. When we validated the rule in patients that were treated later in time and in another hospital, the predictive performance was poor. We show that a simple updating strategy improves the predictive performance in the new patients. Although the example concerns postoperative care, the methodology of the updating method can be applied to all types of prediction rules in any medical domain.

Background

Clinical prediction rules combine patient characteristics (demographic factors, patient history, physical examination, and additional test results) to set a diagnosis or to predict a prognosis. Before prediction rules can be applied in daily practice, their predictive performance needs to be assessed in new patients (validation).[1-3] Unfortunately, while the number of prediction rules in the literature increases, a significantly smaller number is also being validated. This means that for a large number of prediction rules presented in the literature, the predictive performance in new patients is unknown and may be questioned. Further, for a number of clinical outcomes prediction rules have already been developed *and* validated, but some researchers still tend to develop new prediction rules for their own patients. For example, new prediction rules have been developed to predict the presence of pulmonary embolism[4;5], where several prediction rules already existed and were validated.[6-10]

When every new patient sample leads to a new prediction rule, prior information captured in the previous studies and prediction rules is neglected. This is counterintuitive to the notion that research should be based on as many data as possible. The principle of using prior knowledge has been recognised in etiologic and intervention research, in which cumulative meta-analyses are more common. Prior knowledge can also be used in prediction research. First, existing prediction rules should be validated more often. Second, when a rule is indeed validated and it does not show sufficient performance at first sight, researchers should not pursue to develop a new rule from their validation data. A better strategy would be to update the prediction rule, by combining the information of the original prediction rule with the information of the new patients. As a result, the updated rule is based on all available information and the prior information captured in the original rule will not be wasted. This resembles the principle of cumulative meta-analyses.

We present an example in which a simple updating method improved the predictive performance of a prediction rule in the new patients. Although the example concerns postoperative care, the methodology of the updating method can be applied to all types of prediction rules in any medical domain.

Clinical example: a prediction rule for severe postoperative pain

Moderate to severe acute postoperative pain occurs frequently after surgery. Incidences of up to 50% in inpatients and 40% in outpatients (patients that undergo ambulatory surgery) have been reported.[11-14] Risk-based prophylactic treatment could reduce these postoperative pain incidences.

A prediction rule that preoperatively predicts the risk of severe postoperative pain was developed with multivariable logistic regression (article submitted for publication, available on request with corresponding author). The rule was developed on the data of 1944 surgical patients, selected in the Academic Medical Center Amsterdam, the Netherlands (development set).[15]

Severe acute postoperative pain (called ‘postoperative pain’ from now on) was defined as a score ≥ 6 at a numerical rating scale (0 indicates no pain at all, and 10 the most severe pain imaginable), which occurred at least once within the first hour after surgery. The predictors in the rule were gender, age, type of surgery, expected incision size, preoperative pain, preoperative anxiety, need for information and type of patient (inpatient

or outpatient). The prediction rule is presented in Figure 1a in a regression formula (intercept and regression coefficients (β)) and in Figure 1b as an easy to use score chart. The regression formula and the score chart give similar predicted risks. Consider for example in an inpatient setting (intercept = -0.42, corresponding score = 0), a female patient (β = -0.004, score = 0) of age 64 (β = -0.009*64 = -0.576, score = -3), with a preoperative pain score of 7 (β = 0.11*7 = 0.77, score = 4) who is scheduled for a high pain procedure (β = 1.05, score = 5) with a small expected incision size (β = 0, score = 0), who has a preoperative anxiety score of 16 (β = 0.05*16 = 0.8, score = 4) and a preoperative need for information score of 4 (β = -0.05*4 = -0.20, score = -1). The intercept plus the regression coefficients times the predictor values sum op to 1.42, which results in a predicted risk of pain of $1/(1+e^{-1.42}) = 80\%$. The score chart sums up to a total sum score of 9, which results in a risk of postoperative pain of 81%.

We studied the predictive performance of the rule in 1035 new patients (validation set), to test whether the rule is generalisable across time and place. Missing values were imputed with multiple imputation techniques (AregImpute, standard available in S-plus software version 6.2).

As a minimum of 100 events is needed to detect changes in the predictive performance between two sets, the validation set is large enough for this purpose.[16] The patients of the validation set were scheduled for surgery more recently (between February and December 2004) and in a different academic hospital (University Medical Center Utrecht, the Netherlands). One third of the patients in the validation set reported severe pain (36%), compared to 62% of the patients in the development set. The distribution of most predictors was similar in the two datasets, although the patients in the validation set were slightly older (47 versus 43 years), they more often had ambulatory surgery (43% versus 23%) and less often had large expected incision sizes than patients in the development set (7% versus 37%).

Figure 1a Upper part: Original (logistic) regression formula to predict the risk of severe acute postoperative pain for inpatients and outpatients, in which -0.42 is the intercept and the other numbers are the regression coefficients (β) of each predictor or interaction term. Separate regression coefficients for gender and type of surgery were estimated for inpatients and outpatients, as the effect of these predictors differed between these type of patients (interaction terms). The predictors female gender, different types of surgery, expected incision size ≥ 10 cm and ambulatory surgery equal 1 if true and 0 otherwise.

$$\log\left(\frac{\text{risk of pain}}{1 - \text{risk of pain}}\right) = \text{linear predictor} =$$

- 0.42 - 0.004*female gender - 0.009*age + 0.50*low pain surgery + 0.92*moderate pain surgery + 1.05*high pain surgery + 1.72*highest pain surgery + 0.39*expected incision size ≥ 10 cm + 0.11*preoperative pain score + 0.05*APAIS anxiety score - 0.05*APAIS need for information score - 0.70*ambulatory surgery + 0.67*ambulatory surgery*female gender - 0.10*ambulatory surgery*low pain surgery - 0.47*ambulatory surgery*moderate pain surgery - 0.07*ambulatory surgery*high pain surgery - 1.51*ambulatory surgery*highest pain surgery

Risk of pain = $1/(1+e^{-\text{linear predictor}})$

Figure 1b Score chart to predict the risk of severe acute postoperative pain for inpatients and outpatients. The scores are based on the regression coefficients of the prediction rule. For each patient, a sumscore can be calculated by counting the scores that correlate to the characteristics of the patient. The total sumscore can be linked to the individual risk of the patients using the box in the lowest part.

Predictor	**Score per predictor**		
	Inpatient	**Outpatient**	
Constant	0	-4	…
Female gender	0	3	…
Age			…
< 35	-1		
35-57	-2		
58-79	-3		
≥ 80	-4		
Preoperative pain score			…
1 or 2	1		
3 or 4	2		
5 or 6	3		
7 or 8	4		
9	5		
10	6		
Large incision size (≥ 10 cm)	2		…
Type of surgery			…
Lowest pain	0	0	
Low pain	3	2	
Moderate pain	5	2	
High pain	5	5	
Highest pain	9	1	
Preoperative anxiety			…
4, 5 or 6	1		
7, 8, 9 or 10	2		
11, 12, 13 or 14	3		
15, 16, 17 or 18	4		
19 or 20	5		
Preoperative need for information			…
2, 3, 4 or 5	-1		
6, 7, 8 or 9	-2		
10	-3		
			+
Total sum score			…

Total sum score	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12
Risk (%)	19	22	25	29	30	31	33	37	42	49	52	57	61	67	70	74	77	81	83	84	87

Performance and updating of the prediction rule

We considered two aspects of the performance of the prediction rule in the validation set: calibration and discrimination.

Calibration

Calibration is the agreement between the risks as predicted by the prediction rule and the observed frequencies of postoperative pain. Calibration can be graphically assessed with a calibration plot (Figure 2). The dotted line shows the ideal situation in which the predicted risks and the observed frequencies of postoperative pain are completely in agreement. The solid line shows the observed relation between the predicted risks and the observed frequencies. Prediction rules usually show good calibration in the patients that were used to develop the rule. Indeed, the calibration line was very close to the ideal line for the patients in the development set (Figure 2a). However, the prediction rule showed poor calibration when the rule was tested in the patients of the validation set (Figure 2b). The predicted risks of the patients in the validation set were systematically higher than the observed frequencies. This corresponds to the much higher incidence of postoperative pain in the development set (62%) compared to the validation set (36%). The difference in incidence between the two data sets might explain the systematically too high predicted risks in the validation set. However, when the difference in incidence is a result of differences in predictor values, the calibration is not necessarily poor. For example, if the lower incidence in the validation set was a result of a larger proportion of older patients, who experience less often postoperative pain (regression coefficient = - 0.009), or a larger proportion of patients with a low pain surgery type (regression coefficient of 0.50 compared to 1.72 for highest pain surgery), this would have resulted in lower predicted risks. The mean predicted risk would then also be lower and closer to the observed incidence. Accordingly, the calibration plot of the rule in the validation set would then be more similar to figure 2a. However, although the distributions of most predictors were similar in the two datasets, the mean predicted risk was higher than the observed incidence (58% versus 36%). Therefore, the difference in incidences and thus the overestimation of the probabilities by the rule in the validation set should be explained by other characteristics that were not included in the original rule. For example, the Academic Medical Center Amsterdam may have used less aggressive pain treatment than the University Medical Center Utrecht. The effect of characteristics that are not included in the prediction rule is captured in the intercept. Accordingly, a simple updating method that adjusts the intercept of the original prediction rule (while the regression coefficients remain unchanged) can be sufficient to improve the calibration in the present situation. Doing so, the rule is adjusted to the new circumstances and combines the information that is captured in the original rule with the information of the new patients.[17;18] The intercept of the original rule can easily be adjusted such that the mean predicted risk equals the observed incidence in the validation set.[19;20] This is done with a correction factor that is based on the predicted and observed outcome frequency in the validation set. This correction factor equals

$$\ln\left(\frac{\frac{0.361}{1-0.361}}{\frac{0.577}{1-0.577}}\right) = \ln(0.414) = -0.89$$

where 0.361 is the observed incidence in the validation set and 0.577 is the mean predicted risk in the validation set.

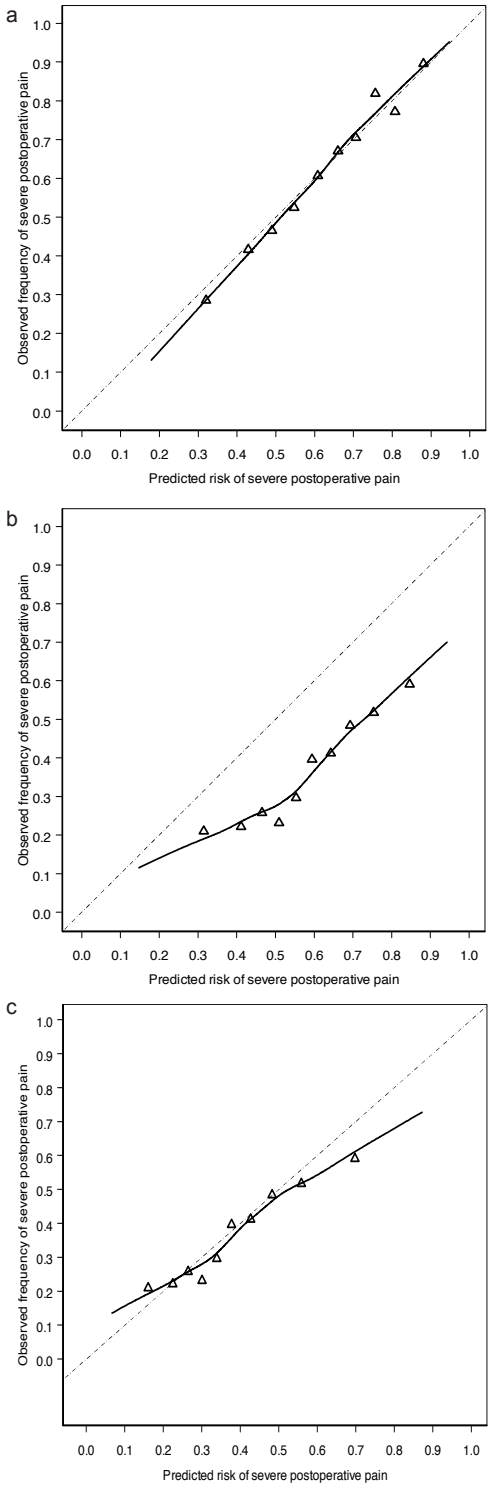


Figure 2 Calibration line of the original prediction rule in the development set (a), in the validation set (b) and the calibration line of the original prediction rule with adjusted intercept in the validation set (c). Triangles indicate the observed frequency of severe acute postoperative pain per decile of predicted risk. The solid line shows the relation between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration, in which the predicted risks equal the observed frequencies of severe postoperative pain.

This factor is the natural logarithm of the odds ratio of the mean observed incidence and the mean predicted risk. The factor simply needs to be added to the intercept of the original rule (upper part of Figure 1), which results in a new intercept of - 0.42 - 0.89 = -1.31.

When we apply the updated rule with the adjusted intercept to the example of the female patient we presented before, the formula would sum up to 0.53 (1.42 – 0.89), which results in a risk of $1/(1+e^{-0.53})$= 63% (versus 80% before updating). In the score chart, the new intercept needs to be multiplied by five, which results in new constant of -7 (-1.31*5= -6.55), resulting in a total sum score of 2, leading to a risk of postoperative pain of 60%. The calibration plot of the updated rule in the validation set is shown in Figure 2c. As expected, the updated prediction rule resulted in lower predicted risks and a calibration line that was much closer to the ideal line.

Discrimination

Discrimination is the ability of the rule to distinguish the patients with postoperative pain from patients without postoperative pain, and is quantified with the area under the ROC curve (AUC). An AUC ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination). The AUC of the original prediction rule (before updating) in the validation set was 0.65 (0.57 - 0.73), compared to 0.71 (0.66 - 0.76) in the development set. As expected, the updating strategy did not change the discriminative ability. The adjustment of the intercept does not alter the ranking of the predicted risks of the individual patients, and since the AUC is a rank order statistic, the AUC of the updated rule was also 0.65 in the validation set.

Conclusion

In a situation of poor performance in new patients, many researchers are tempted to develop their own prediction rule for the same problem at hand, by re-estimating the regression coefficient of all predictors with the new data, or even by selecting the predictors again. When the collection of every new patient sample results in a new prediction rule, all the prior information that is captured in the previous studies and prediction rules is neglected. This is counterintuitive to the notion that (medical) research results are more precise when they are based on data of more patients. Therefore, prediction rules should also be based on data of as many patients as possible.[21] A valuable and simple alternative to the re-development of prediction rules is to update the previously developed prediction rules – provided that they were carefully developed – with the data of the new patients. Updating methods for prediction rules can vary in the degree of adjustment.[18] A comprehensive discussion of all available methods for updating goes beyond the scope of this paper. However, often a simple adjustment such as the re-estimation of the intercept as we illustrated already improves the performance of a prediction rule in new patients enormously.[22] Moreover, such updating will likely improve the generalisibility of the updated rule across new populations because the rule is based on more data. With these advances, the future may be one in which prediction rules are continuously validated and (if necessary) updated, maintaining all available information.

References

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

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

3 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201-9.

4 Chen JY, Chao TH, Guo YL, Hsu CH, Huang YY, Chen JH et al. A simplified clinical model to predict pulmonary embolism in patients with acute dyspnea. Int Heart J. 2006;47:259-71.

5 Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. Ann Intern Med. 2006;144:165-71.

6 Tamariz LJ, Eng J, Segal JB, Krishnan JA, Bolger DT, Streiff MB et al. Usefulness of clinical prediction rules for the diagnosis of venous thromboembolism: a systematic review. Am J Med. 2004;117:676-84.

7 Righini M, Bounameaux H. External validation and comparison of recently described prediction rules for suspected pulmonary embolism. Curr Opin Pulm Med. 2004;10:345-49.

8 Perrier A, Roy PM, Aujesky D, Chagnon I, Howarth N, Gourdier AL et al. Diagnosing pulmonary embolism in outpatients with clinical assessment, D-dimer measurement, venous ultrasound, and helical computed tomography: a multicenter management study. Am J Med. 2004;116:291-99.

9 Chunilal SD, Eikelboom JW, Attia J, Miniati M, Panju AA, Simel DL et al. Does this patient have pulmonary embolism? JAMA. 2003;290:2849-58.

10 Chagnon I, Bounameaux H, Aujesky D, Roy PM, Gourdier AL, Cornuz J et al. Comparison of two clinical prediction rules and implicit assessment among patients with suspected pulmonary embolism. Am J Med. 2002;113:269-75.

11 Chauvin M. State of the art of pain treatment following ambulatory surgery. Eur J Anaesthesiol. 2003;20:3-6.

12 Huang N, Cunningham F, Laurito CE, Chen C. Can we do better with postoperative pain management? Am J Surg. 2001;182:440-448.

13 Shaikh S, Chung F, Imarengiaye C, Yung D, Bernstein M. Pain, nausea, vomiting and ocular complications delay discharge following ambulatory microdiscectomy. Can J Anaesth. 2003;50:514-18.

14 Svensson I, Sjostrom B, Haljamae H. Assessment of pain experiences after elective surgery. J Pain Symptom Manage. 2000;20:193-201.

15 Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea with vomiting and economic analysis. Anesthesiology. 2001;95:616-26.

16 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475-83.

17 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation. 1999;99:2098-104.

18 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage.

Stat Med. 2004;23:2567-86.
19 Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med. 1986;105:586-91.
20 Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med. 1986;146:81-83.
21 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.
22 van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med. 2000;19:3401-15.

Chapter 3.4

Updating methods improved the performance of a clinical prediction model in new patients

Abstract

Objective Ideally, clinical prediction models are generalisable to other patient groups. Unfortunately, they perform regularly worse when validated in new patients and are then often redeveloped. While the original prediction model usually has been developed on a large dataset, redevelopment then often occurs on the smaller validation set. Recently, methods to update existing prediction models with the data of new patients have been proposed. We used an existing model that preoperatively predicts the risk of severe postoperative pain (SPP) to compare five updating methods.

Study design and Setting The model was tested and updated with a set of 752 new patients (274 (36%) with SPP). We studied the discrimination (ability to distinguish between patients with and without SPP) and calibration (agreement between the predicted risks and observed frequencies of SPP) of the five updated models in 283 other patients (100 (35%) with SPP).

Results Simple recalibration methods improved the calibration to a similar extent as revision methods that made more extensive adjustments to the original model. Discrimination could not be improved by any of the methods.

Conclusion When the performance is poor in new patients, updating methods can be applied to adjust the model, rather than developing a new model.

Introduction

Setting a diagnosis or a prognosis in clinical practice is a multivariable process. Physicians combine patient characteristics and test results to estimate the probability that a disease or outcome is present (diagnosis) or will occur (prognosis). To guide physicians in their estimation of diagnostic and prognostic probabilities, many clinical prediction models have been developed. These prediction models are algorithms that combine patient characteristics to set a diagnosis or to predict a prognosis. Such models can improve the identification of high risk patients, and guide medical decision making. Besides that, they can be used as a quality of care instrument, since the predicted risks can be compared with the observed actual outcomes. The performance of prediction models needs to be tested in new patients before the models can be applied in clinical practice with confidence.[1;2] Unfortunately, the predictive performance is often decreased when a model is tested in new patients, compared to the performance estimated in the patients that were used to derive the model. As a consequence, the original prediction model is frequently rejected and a new prediction model is developed. While the original prediction model usually has been developed on a large dataset, redevelopment then often occurs on the much smaller dataset of the new patients only. When every new patient sample leads to a new prediction model, the prior information that is captured in previous studies and prediction models is neglected. This is counterintuitive to the notion that research should be based on data of as many patients as possible. The principle of using knowledge of previous studies has been recognized in etiologic and intervention research, in which cumulative meta-analyses are more common. The alternative to re-developing prediction models in every new patient sample is updating the existing prediction models. The updated models combine the information that is captured in the original model with the information of the new patients.[3-7] As a result, the updated models are adjusted to the new patients.

Recently, several updating methods have been proposed in the statistical literature.[3] The methods vary in extensiveness, which is reflected by the number of parameters that is adjusted or reestimated. A relatively simple recalibration method, for instance, results in an updated model that has a new intercept and adjusted regression coefficients that are based on multiplication of the original coefficients with a single recalibration factor. With this method, only two parameters are estimated. More extensive updating methods estimate more parameters, for instance when all individual regression coefficients are reestimated.

In this study, we present an empirical example of updating a model to predict the risk of severe postoperative pain. We show the results of simple and more extensive updating methods.

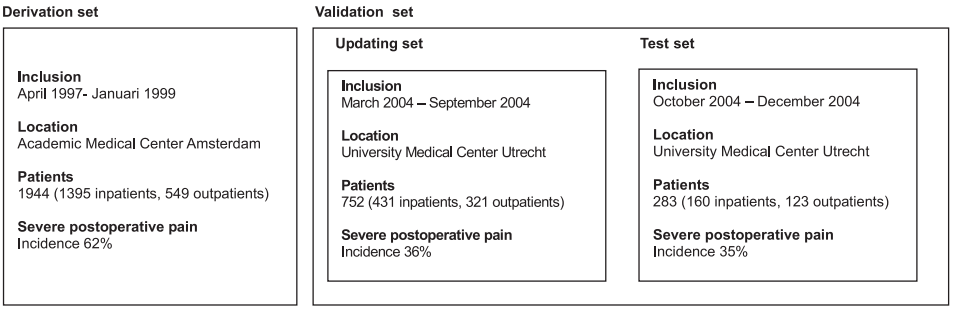
Patients and methods

Severe postoperative pain occurs frequently after surgery. Patients at high risk of severe postoperative pain may benefit from preventative strategies. Therefore, a prediction model has been developed to predict the risk of acute severe postoperative pain.

The datasets

Usually, prediction models are developed in derivation sets and tested in validation sets. However, since we wanted to show the effect of the updating methods, the validation set was split into an updating set and a test set (in which the updated model is tested). As a result, three datasets were used in this study (Figure 1).

Figure 1 Schematic presentation of the characteristics of the derivation, updating and test dataset



The derivation set

The dataset that was used to derive the prediction model for severe postoperative pain in surgical patients included 1944 surgical patients (1395 inpatients and 549 outpatients).[8] In brief, the patients, aged 18-85, were selected from a randomized trial to investigate whether intravenous anaesthesia (propofol) yielded a lower incidence of postoperative nausea and vomiting than inhalation anaesthesia (isoflurane and nitrous oxide).[9] The trial was conducted at the Academic Medical Center of the University of Amsterdam, the Netherlands between April 1997 and January 1999. All types of surgery were included, except cardiac surgery and intracranial neurosurgical procedures. Exclusion criteria were emergency surgery, pregnancy, ASA physical status 4 and morbid obesity (weight > 120 kg).

ASA physical status denotes the American Society of Anesthesiologists preoperative risk classification based on comorbidity, and ranges from 1 (healthy patient) to 4 (patient with comorbidity that is a constant threat to life). Besides postoperative nausea and vomiting, the severity of acute postoperative pain was recorded with a numerical rating scale (NRS) score (where 0 indicates no pain at all, and 10 the most severe pain imaginable) every 15 minutes after arrival at the Post Anaesthesia Care Unit (PACU) for one hour (four measurements). Severe postoperative pain was defined as a NRS score of 6 or higher on at least one of the four measurement points.

The updating and test set

The updating and test set together contain 1035 adult surgical patients (591 inpatients and 444 outpatients) scheduled for surgery in 2004 in a second hospital, the University Medical Center of Utrecht, the Netherlands (Figure 1). Similar as in the derivation set, all types of surgery were included, except emergency surgery, cardiac surgery and intracranial neurosurgical procedures. Also, the same methods of data collection were applied regarding the predictors. However, a small change was made to the assessment of the outcome for logistic reasons. The NRS of the inpatients was scored 15 and 60 minutes after arrival at the PACU. The NRS of the outpatients was scored at arrival and after 15, 30 and 45 minutes on the PACU.

A test set should include at least 100 events to contain reasonable power to detect changes in predictive performance of prediction models between two datasets.[10] We therefore included 283 patients in the test set, of which 100 patients had severe postoperative pain and 183 patients had no severe postoperative pain (outcome incidence 35%). To comply with the chronology in practice, we did not randomly select these 283 patients in the test set, but rather the more recently treated patients.

The prediction model

The prediction model predicts the risk of severe acute postoperative pain in the first hour after surgery. The model was developed by multivariable logistic regression and included eight predictors that were assessed during the preoperative visit within a few weeks before surgery: gender, age, type of surgery, expected incision size, preoperative pain, anxiety score, need for information score and type of patient. The classification of type of surgery (lowest pain, low pain, moderate pain, high pain and highest pain) was developed in the derivation study, based on clinical experience of surgeons and anesthesiologists, and the associations we found in the data.[8] The expected incision size of the surgical procedure was dichotomised into smaller than 10 cm and equal to or larger than 10 cm. The severity of preoperative pain was assessed with the pain domain of the quality of life questionnaire Short Form 36 (SF-36), that was transformed to a score of 0 (no pain) to 10.[11] The anxiety score and need for information score were assessed with the Amsterdam Preoperative Anxiety and Information Scale (APAIS). The APAIS consists of six questions, each scored on a 5-point Likert scale from 1 (not at all) to 5 (extremely). Four questions are used to assess the patients' preoperative anxiety score (range 4-20) and two questions are used to assess the patients' need for information regarding the scheduled surgery and anesthesia (range 2-10).[12] Type of patient was defined as inpatient or outpatient. Interaction terms were included since the effect of some predictors was different for inpatients and outpatients.[8]

Most predictors were completely observed for all patients, except incision size (1 value missing, 0.07% of patients), preoperative pain (38 values missing, 2.7% of patients), anxiety (27 values missing, 1.9% of patients) and need for information (34 values missing, 2.4% of patients).

The formula of the prediction model is of the form:

$$\log\left(\frac{\text{risk of pain}}{1-\text{risk of pain}}\right) = \text{linear predictor} = ß_0 + ß_1*\text{predictor}_1 + ... + ß_n*\text{predictor}_n \quad \text{(Formula 1)}$$

In this formula, $ß_0$ is the intercept and $ß_1$ till $ß_n$ are the regression coefficients, which are presented in Table 1.

The risk of severe postoperative pain in individual patients (scale: 0-100%) can be calculated with the formula:

$$\text{risk} = \frac{1}{1 + e^{-\text{linear predictor}}}$$

The model was also presented as an easy to use score chart.[8]

The area under the Reciever Operating Characteristic (ROC) curve (AUC) of the model was 0.71 in the derivation set.

Analyses

Missing values

Deleting subjects with missing values often leads to bias and a loss of statistical power.[13-15] All missing values were imputed with multiple imputation techniques using all other available information of the patients (aregImpute in the Design Library, standard available in S-plus software version 6.2). Logistic regression models were used to impute dichotomous missing values, polytomous logistic regression models for categorical missing values and linear regression models for continuous missing values.

Table 1 Intercept and regression coefficients of the prediction model for severe postoperative pain

Predictor	Regression coefficient
Intercept	-0.42
Female gender	-0.004
Age	-0.009
Type of surgery	
Lowest expected pain (reference)	-
Low expected pain	0.50
Moderate expected pain	0.92
High expected pain	1.05
Highest expected pain	1.72
Expected incision size ≥ 10 cm	0.39
Preoperative pain score	0.11
Anxiety score	0.05
Need for information score	-0.05
Ambulatory surgery	-0.70
Ambulatory surgery * female gender	0.67
Ambulatory surgery * type of surgery	
Ambulatory surgery * lowest expected pain (reference)	-
Ambulatory surgery * low expected pain	-0.10
Ambulatory surgery * moderate expected pain	-0.47
Ambulatory surgery * high expected pain	-0.07
Ambulatory surgery * highest expected pain	-1.51

Predictive performance

To study the performance of the prediction model in the updating and test set, we assessed its calibration and discrimination. Calibration refers to the agreement between the predicted risks and observed frequencies of severe postoperative pain. This was graphically assessed with a calibration plot that shows a calibration line, which can be described with a slope and an intercept (calibration slope and calibration intercept).[16;17] These parameters were first estimated in the updating set using a logistic regression model with the presence of severe postoperative pain (yes/no) as the outcome variable and the linear predictor of the original prediction model (Formula 1) as the only covariate. The calibration slope and calibration intercept are ideally 1 and 0 respectively (perfect calibration). Since a calibration intercept is difficult to interpret when the calibration slope is not equal to 1, the calibration intercept is estimated with the calibration slope fixed at 1. This is done by fitting the logistic regression model with the regression coefficient of the linear predictor fixed at 1. A calibration intercept different from 0 indicates that the model's predicted probabilities in the updating set are systematically too high (intercept < 0) or too low (intercept > 0). Such a difference in calibration intercept commonly reflects a difference in outcome incidence between the derivation set and the updating set that can not be ex-

plained by different distributions of the predictor values. A calibration slope smaller than 1 indicates optimism; the regression coefficients of the original model were too large, which results in too extreme predictions in the new patients (low predictions are too low and high predictions are too high). A calibration slope that is larger than 1 indicates that the regression coefficients of the original model were too close to zero.

Discrimination is the ability of the model to discriminate patients with severe pain from patients without severe pain. It was quantified with the area under the ROC curve (AUC), which is equal to the c-statistic for a dichotomous outcome variable.[18]

Updating methods

Several methods can be used to update a prediction model when it shows poor performance in new patients. The methods we used are shown in Table 2 and have been presented previously in the statistical literature.[3]

Table 2 Updating methods for the prediction rule for severe acute postoperative pain. Method 1 and 2 are recalibration methods, while method 3, 4 and 5 are revision methods

No.	Updating method
0	No adjustment (the original prediction rule)
1	Adjustment of the intercept using the calibration intercept
2	Adjustment of the intercept and the regression coefficients using the calibration intercept and calibration slope
3	Extra adjustment of predictors with a different effect in the updating set compared to the derivation set, after recalibration by method 2
4	Reestimation of the intercept and the regression coefficients of all predictors, using the data of the updating set
5	Reestimation of the intercept and the regression coefficients of all predictors, using the combined data of the derivation and updating set

In method 0, no adjustments are made to the original prediction model (Formula 1). Method 1 and method 2 are so-called 'recalibration' methods. Method 1 intends to correct the 'calibration in the large', such that the mean predicted probability is equal to the observed outcome frequency.[3] Only the intercept of the original model is adjusted.[4;5] This can be done by fitting a logistic regression model with the linear predictor (as offset) as the only covariate in the updating set or by calculating a correction factor (Formula 2) that is based on the mean predicted risk and observed outcome frequency.

$$\text{correction factor} = \ln\left(\frac{\frac{\text{observed outcome frequency}}{1-\text{observed outcome frequency}}}{\frac{\text{mean predicted risk}}{1-\text{mean predicted risk}}}\right) \qquad \text{(Formula 2)}$$

This correction factor equals the calibration intercept when the outcome frequency is not extremely low or high. The correction factor simply needs to be added to the intercept of the original prediction model, which results in a new intercept.

In method 2, both the calibration intercept and calibration slope are estimated with Formula 3 to adjust the original model, also called 'logistic calibration'.[13] A logistic regression model is fitted with the linear predictor as the only covariate in the updating set.

$$\ln\left(\frac{\text{risk of pain}}{1-\text{risk of pain}}\right) = \alpha_{\text{calibration}} + \beta_{\text{calibration}}\text{*linear predictor} \qquad \text{(Formula 3)}$$

The calibration slope $\beta_{\text{calibration}}$ is used to recalibrate the original regression coefficients, i.e. the regression coefficients of the original model are multiplied with the calibration slope. A calibration slope equal to 1 means that the original regression coefficients do not need adjustment. The intercept of the original prediction model is adjusted by adding the calibration intercept $\alpha_{\text{calibration}}$.

Updating methods 3, 4 and 5 are so-called revision methods. Method 3 builds on method 2. It tests whether the effect of the individual predictors is different in the updating set compared to the derivation set, after the logistic calibration of method 2.

$$\ln\left(\frac{\text{risk of pain}}{1-\text{risk of pain}}\right) = \alpha_{\text{calibration}} + \beta_{\text{calibration}}\text{*linear predictor} + \gamma\text{*predictor} \qquad \text{(Formula 4)}$$

The logistic regression model of Formula 4 is first fitted for each predictor of the original model separately. The γ is the deviation from the recalibrated regression coefficient (based on method 2). When γ is unequal to 0, the effect of the predictor is still different in the updating set, after the recalibration of method 2. We test with the likelihood ratio test ($p < 0.05$) whether the deviation has added predictive value. We selected the predictors with forward selection, starting with the predictor with the largest Wald statistic. We note that a 'predictor' in Formula 4 can also be an interaction term.

In method 4 and 5, the intercept and the regression coefficients of all predictors in the original model are reestimated, without variable selection. In method 4, only the updating set was used and in method 5 the combined derivation and updating set. Since the incidence in the derivation set was quite different from the updating and test set (62%, 36% and 35% respectively), the incidence in the combined derivation and updating set was higher than in the test set. Therefore, for method 5, the intercept of the updated model was first adjusted by applying updating method 1.

Shrinkage methods

Newly estimated individual regression coefficients (such as in methods 3, 4 and 5) are often too optimistic, due to overfitting. This can be overcome by shrinkage of the regression coefficients. Regression coefficients can be shrunk either towards zero[3;19] or towards the recalibrated values (the regression coefficients of the model updated with method 2). Shrinkage towards zero needs to be applied when no information is present about the strength of the effect of the predictors in the model.[19] For method 5, where we combined the derivation set and the updating set and ignored the regression coefficients of the original model, all regression coefficients were in fact estimated without prior knowledge. As a result, the regression coefficients of the model updated with method 5 were shrunk towards zero using the following formula

$$\beta_{\text{method5}} = \text{shrinkage factor}\text{*}\beta_{\text{reestimated}} \qquad \text{(Formula 5)}$$

In this formula, the shrinkage factor is a heuristic shrinkage factor[19], calculated as

$$\text{shrinkage factor} = \frac{\chi^2_{\text{updated-nullmodel}} - df}{\chi^2_{\text{updated-nullmodel}}} \qquad \text{(Formula 6)}$$

in which $\chi^2_{\text{updated-nullmodel}}$ is the difference in -2loglikelihood of a model with only an intercept and the updated model, and df equals the difference in degrees of freedom between these models.

Shrinkage towards the recalibrated values can be applied when there is prior knowledge about the regression coefficients. In method 3, the same predictors and regression coefficients were used as in the original model. The only difference was that the deviation of the predictors in the derivation and updating set was modelled. Therefore, the deviations of the regression coefficients of the model updated with method 3 are shrunk towards the recalibrated values.[3] The regression coefficients of method 3 were shrunk with the formula

$$\beta_{\text{method3}} = \beta_{\text{recalibrated}} + \text{shrinkage factor}*\gamma \qquad \text{(Formula 7)}$$

In this formula, $\beta_{\text{recalibrated}}$ is the recalibrated regression coefficient (based on method 2) and γ is the deviation after updating with method 3 from the recalibrated coefficient. The shrinkage factor can be calculated as

$$\text{shrinkage factor} = \frac{\chi^2_{\text{updated-recalibrated}} - df}{\chi^2_{\text{updated-recalibrated}}} \qquad \text{(Formula 8)}$$

were $\chi^2_{\text{updated-recalibrated}}$ is the difference in – 2LogLikelihood between a model with estimated deviations of individual regression coefficients and the recalibrated model, and df stands for the difference in degrees of freedom between these models.

Also for method 4, shrinkage towards the recalibrated values could be applied as there was prior knowledge about the regression coefficients. The same predictors were used as in the original model, only the regression coefficients were reestimated. Hence, the regression coefficients of method 4 were shrunk with the formula

$$\beta_{\text{method4}} = \beta_{\text{recalibrated}} + \text{shrinkage factor}*(\beta_{\text{reestimated}} - \beta_{\text{recalibrated}}) \qquad \text{(Formula 9)}$$

In this formula, $\beta_{\text{reestimated}}$ is the reestimated regression coefficient. The shrinkage factor can be calculated from Formula 8.

Results

Approximately one third of the patients in the updating (274, 36%) and test set (100, 35%) reported severe postoperative pain in the first hour after surgery, compared to 62% (1205) of the patients in the derivation set (Table 3). Patients in the updating and test set were slightly older, they underwent more often ambulatory surgery and had less often a large expected incision size than patients in the derivation set. The distribution of patient characteristics in the updating and test set were very similar, although the test set included more women (183, 65%) than the updating set (413, 55%), which must be due to chance.

Table 3 Distribution of the predictors in the derivation set, the updating set and the test set, n (%) unless stated otherwise.

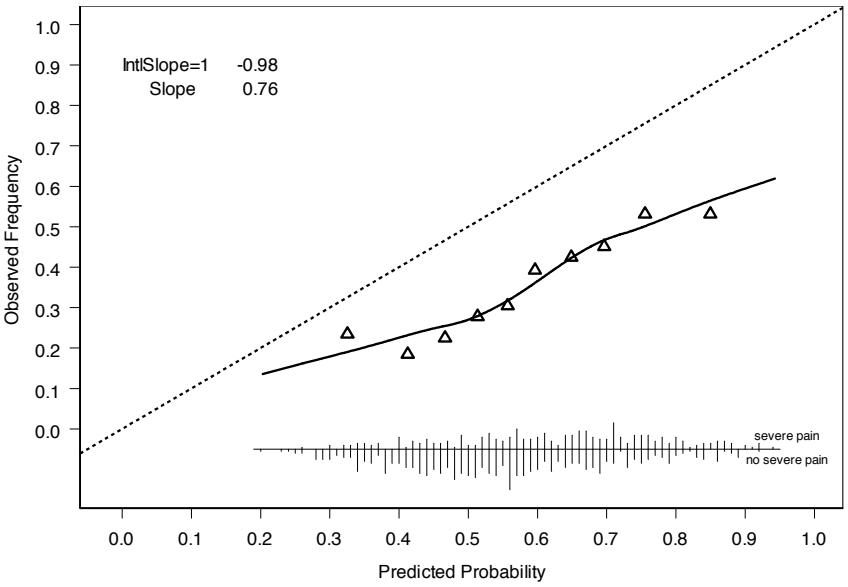
Patient characteristics	Derivation set (n=1944)		Updating set (n=752)		Test set (n=283)	
Female gender	1110	(57%)	413	(55%)	183	(65%)
Age (years)[a]	43	(15)	46	(16)	48	(16)
Ambulatory surgery	549	(28%)	321	(43%)	123	(43%)
Preoperative pain score[a]	3.0	(2.7)	3.4	(2.9)	3.3	(2.9)
Type of surgery						
Lowest pain	72	(4%)	61	(8%)	31	(11%)
Low pain	590	(30%)	216	(29%)	81	(29%)
Moderate pain	348	(18%)	139	(18%)	61	(22%)
High pain	671	(35%)	252	(34%)	87	(31%)
Highest pain	263	(14%)	84	(11%)	23	(8%)
Expected incision size ≥ 10 cm	724	(37%)	50	(7%)	17	(6%)
APAIS anxiety score[a]	9.4	(4.0)	9.5	(3.8)	9.7	(4.0)
APAIS need for information[a]	6.5	(2.2)	6.2	(2.3)	6.2	(2.4)
Severe acute postoperative pain, NRS ≥ 6	1205	(62%)	274	(36%)	100	(35%)

[a] Mean (standard deviation)

Predictive performance of the original prediction model (method 0)

The calibration plot showed that the predicted probabilities of the original prediction model were systematically too high (bias) in the updating set, as indicated by the solid line (Figure 2). The calibration slope was 0.76, indicating that the originally estimated regression coefficients were too large, resulting in too extreme predictions in the new patients. The calibration intercept given slope=1was -0.98, indicating that the predicted probabilities were systematically too high. The discriminative ability of the model was lower in the updating set (AUC = 0.65 (0.57 - 0.73)), than in the derivation set (0.71 (0.66-0.76)). Given the poor calibration and lower AUC, an updated prediction model may have better predictive performance in new patients.

Figure 2 Calibration plot of the original prediction model in the updating set (Updating method 0).



The updated prediction models

Table 4 shows the estimates of the main parameters of the various updating methods. The regression coefficients of all predictions per updated model are presented in the Appendix. In method 1, only the model intercept was updated. Corresponding with the too high predicted risks in the updating set, the intercept of the updated model was decreased with -0.98, in accordance with the estimated calibration intercept described above. In method 2 (adjusting the intercept and regression coefficients by the calibration intercept and calibration slope respectively), the individual regression coefficients were multiplied with 0.76 and the intercept of the original prediction model was decreased with - 0.77. This implicates that although the calibration slope corrects the original regression coefficients (that were too large), an adjustment in the intercept was still needed. However, the adjustment to the intercept of the original model was smaller (-0.77) than the adjustment that was needed without the adjustment of the regression coefficients with the calibration slope (-0.98 in method 1). The updated regression coefficients are calculated by multiplying the calibration slope of method 2 (0.76) with the regression coefficient of the original model. For ambulatory surgery, for instance, this results in a regression coefficient of 0.76*(-0.70) = -0.53.

Method 3 (revision of the predictors with a different association in the updating set than in the derivation set) showed a statistically significant different effect of the interaction between ambulatory surgery (outpatients versus inpatients) and type of surgery between the updating set and the derivation set ($p < 0.001$). The effects of the other predictors were not significantly different. The updated regression coefficient of the interaction between ambulatory surgery and type of surgery, for instance the interaction between low pain procedure and ambulatory surgery, was obtained by multiplying the calibration slope (0.90) with the regression coefficient of the original model (-0.10).

Table 4 Estimated parameters of the updating methods 1, 2 and 3 (see text for more information about the updating methods). The reestimated parameters of methods 4 and 5 are not presented in this table. The (reestimated) regression coefficients of all updated models are presented in the Appendix.

	Method 1	Method 2	Method 3	Method 4	Method 5
Calibration intercept	-0.98[a]	-0.77	-1.01	-	-
Calibration slope	-	0.76	0.90	-	-
Deviation from recalibrated regression coefficients[b]					
Ambulatory surgery * low pain	-	-	-0.54	-	-
Ambulatory surgery * moderate pain	-	-	0.99	-	-
Ambulatory surgery * high pain	-	-	0.35	-	-
Ambulatory surgery * highest pain	-	-	2.65	-	-
Shrinkage factor	-	-	0.89	0.71	0.96

[a] Calibration intercept with calibration slope fixed at 1
[b] Before shrinkage

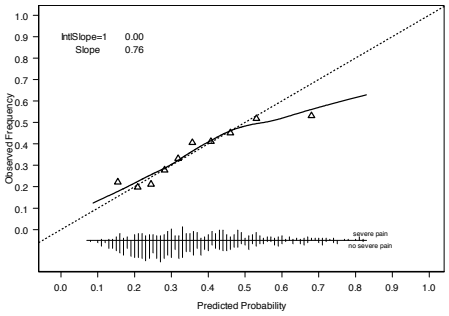
The deviation from this regression coefficient (-0.54, Table 4) was then added, after shrinkage with the shrinkage factor (0.89, Table 4). This resulted in 0.90*(-0.10)+0.89*(-0.54) = -0.57 as the updated regression coefficient of the interaction between low pain procedure and ambulatory surgery (see Appendix). The regression coefficient of the interaction between low pain procedure and ambulatory surgery was more negative than in the derivation set (deviation -0.47). Outpatients that are scheduled for low pain procedures thus had a lower risk of severe postoperative pain in the updating set than outpatients in the derivation set. The deviation of the other types of surgery was in an opposite direction, which indicates that for outpatients that were scheduled for procedures with moderate, high or highest pain, the risk of severe postoperative pain was higher for patients in the updating set than in the derivation set.

Predictive performance of the updated prediction models in the updating and test set
All updated models resulted in improved calibration in the updating set, compared to the original model (Figure 3, left part) Since this set was also used to update the model, intercept and slope were always equal to 0 and 1 by definition (except for the slope of the model updated with method 1 as here only the model's intercept was updated). Methods 1 to 4 showed all good calibration in the large (mean predicted risk equals observed incidence) when the models were tested in the test set. This is reflected by a calibration intercept close to zero (Figure 3, right part). This also applies to method 5, though only after extra adjustment of the intercept. The calibration slopes showed more variability between the updated models in the test set. In method 1, only the calibration intercept was used to update the model and the updated model showed a calibration slope of 0.84. Method 2 resulted in a calibration slope of 1.07 (close to 1), whereas adjustment of the individual regression coefficients with method 3 to 5 did not result in better calibration slopes (0.77, 0.80 and 0.83 respectively).

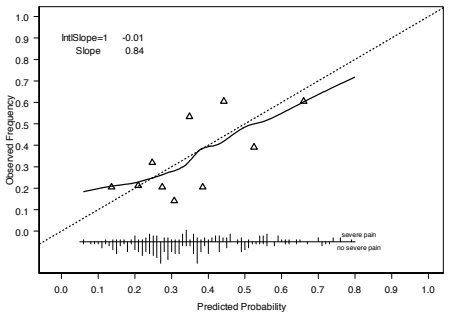
Figure 3 Calibration plots of the prediction model that was updated by method 1,2,3,4 and 5. Calibration plots are shown for the updated models in the updating set and the test set. The updating methods are described in the text. For method 5, the derivation and updating set were combined. Since the incidence of severe postoperative pain in the combined set is higher than in the test set, also the calibration plot of method 5 with an adjusted intercept is presented.

Method 1

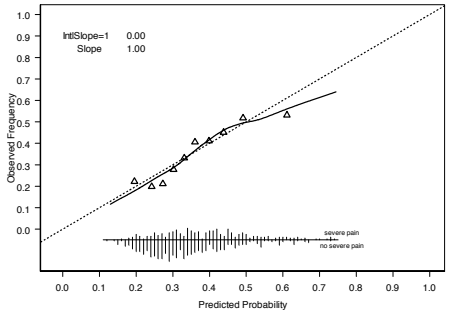
Updating set (n=752)

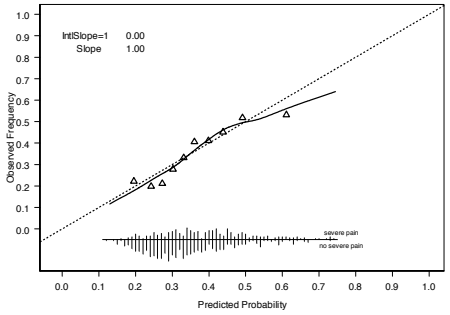


Test set (n=283)

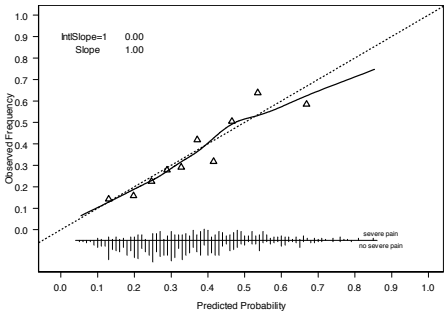


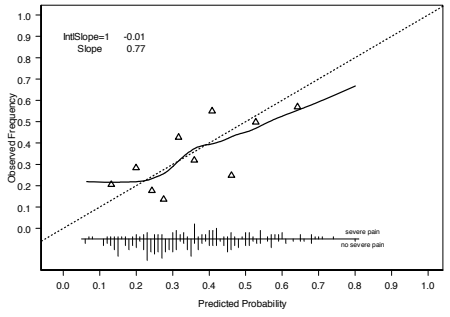
Method 2



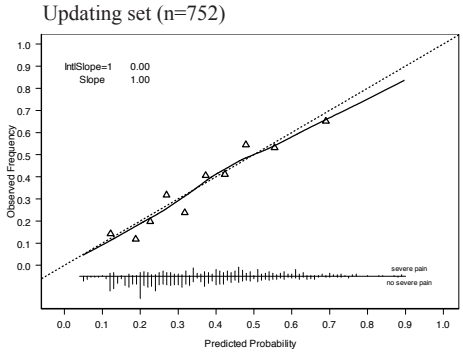


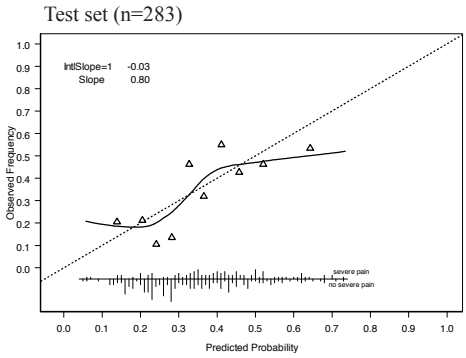
Method 3



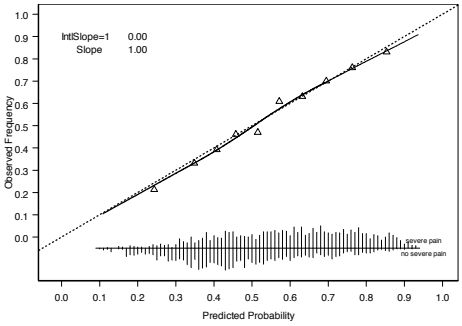


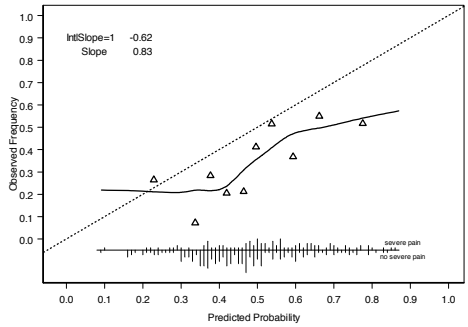
Method 4





Method 5





Adjusted intercept

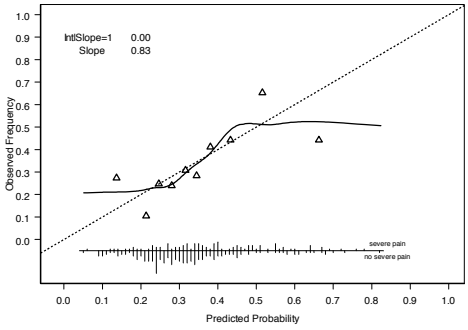


Table 5 shows that the models updated with recalibration method 1 and 2 obviously had the same AUC as the original prediction model in the updating set (0.65). These updating methods do not change the ranking of the predicted risks of patients and thus do not affect a model's discriminative ability. Updating methods 3, 4 and 5 resulted in an increased AUC in the updating set (0.70, 0.71 and 0.72 respectively). However, when tested in the test set, this increase in AUC did not hold and the updated models showed a similar AUC as the original prediction model in the test set (0.66).

Table 5 Discrimination of the updated prediction models in the updating set and test set, expressed by the Area Under the Recieved Operating Characteristic Curve (with 95% Confidence Interval)

	Method 0*	Method 1	Method 2	Method 3	Method 4	Method 5
Updating set (n=752)[a]	0.65	0.65	0.65	0.70	0.71	0.72
	(0.57-0.73)	(0.57-0.73)	(0.57-0.73)	(0.62-0.77)	(0.63-0.79)	(0.68-0.75)
Test set (n=283)	0.66	0.66	0.66	0.66	0.67	0.67
	(0.52-0.80)	(0.52-0.80)	(0.52-0.80)	(0.52-0.79)	(0.54-0.80)	(0.53-0.80)

[a] For method 5, the derivation and updating set are combined into a dataset containing 2696 patients
* Original prediction rule

Discussion

We tested a prediction model for the risk of acute severe postoperative pain in surgical patients. The patients were recruited in a later period and in another hospital than the patients on which the model was developed. As the predictive performance of the model was poor, updating was necessary to adjust the model to the new situation. We compared five updating methods that differed in extensiveness of the updating. In this example study, simple recalibration methods improved the calibration to a similar extent as the revision methods that made more extensive adjustments to the original model. Discrimination was improved in the updating set when the more extensive updating methods were applied, but this improvement did not hold in the test set. We compared updating methods in one empirical dataset. Other datasets can show different results. In general, when the discrimination of a model is sufficient, recalibration methods can improve the calibration. On the other hand, when the discrimination of a model needs to be improved, revision methods are necessary. However, the aim of our example study was not merely to present the best method, but above all to increase the familiarity and appreciation of updating methods of prediction models.

The predicted risks in new patients can be biased, i.e. systematically over- or underestimated.[3;20] This can for instance be caused by variables or characteristics that are not included in the model, but do have an effect on the originally estimated regression parameters (intercept and regression coefficients). Although researches in such situations often pursue to completely reestimate all regression coefficients using the data of the new patients only (method 4), we showed that this did not improve the performance of the model more than the simple recalibration methods (methods 1 and 2) in our example study. Moreover, when researchers expect that a particular predictor in the original model indeed has a different effect in the updating set, revision methods with small adjustments for that predictor (method 3) are still preferred over methods that reestimate all regression coefficients with only the new data (method 4). Another option when reestimating the regression coefficients is to combine the derivation and updating set, if one has indeed access to the derivation data (method 5). However, in our example study, this did also not result in a better performance in the test set than the simple recalibration methods (methods 1 and 2). Our results support the view that researchers should opt first for adjusting the existing model to the new patients by estimating only a few parameters (method 1 and 2, and perhaps 3) rather than directly reestimating or rebuilding a complete new model (method 4 and 5).[3;20]

Choosing between simple and more extensive updating methods is a choice between variance and bias, in which the size of both the derivation and the updating set should be considered. Estimates of regression parameters of prediction models derived on large derivation sets are obviously very precise (low variance). However, as said, the corresponding predicted risks may still be biased in new patients, mainly due to different associations of the predictors in the new patients or to the presence of important predictors that were not included in the model. In case of the latter, adjustment of the model intercept is often sufficient and preferable, as such recalibration removes bias of the original intercept while precision of the regression coefficients is sustained.[3;20] In our example, a simple adjustment of the intercept of the prediction model also resulted in deletion of a substantial part of the bias of the predicted risks, without loosing precision of the other regression coefficients. A few methodological issues need to be addressed. First, there was a large difference in incidence of severe postoperative pain (defined as an NRS score ≥ 6 at least once in the first hour after surgery) between the derivation set and the updating and test sets (62%, 36% and 35% respectively). Since the distribution of the predictor values was very similar between the datasets, the difference in incidence should be explained by characteristics that were not included in the prediction model. One explanation could be the different time points at which the NRS pain scores were assessed. In the derivation set, pain was scored every 15 minutes after arrival at the PACU (15, 30, 45, 60 minutes). In the updating and test set, pain of the inpatients was scored at 15 and 60 minutes, and of the outpatients at 15, 30 and 45 minutes after arrival at the PACU. The lower number of observations may have led to unobserved cases of severe postoperative pain, and therefore to a lower observed incidence of severe postoperative pain. However, we estimated what the incidence of severe postoperative pain would be in the derivation set if the measurements were done on the same time points as in the updating set and test set. For the inpatients, the incidence of severe postoperative pain was 56% (compared to 60%) and for the outpatients 43% (compared to 44%). Hence, the lower number of pain measurements could only to a minor extent have led to the lower incidence of severe postoperative pain in the updating and test set. A more plausible explanation is the difference or change in characteristics that were not included in the original model but do influence the originally estimated associations of the predictors in the model, for example, the change of pain management over time. It may well be that in the later period more aggressive preventative pain management was introduced in daily practice, caused by an increased focus on postoperative outcomes (due to new pain guidelines issued in 2003).

Second, the classification of type of surgery (five groups) was developed in the derivation study. Starting with 27 groups of surgical procedures (based on clinical experience, current practice and interviews with surgeons and anesthesiologists), the univariable association between each surgical group and severe acute postoperative pain was estimated. Groups with similar associations were then combined, resulting in the five categories of surgery. The classification was thus partly data-driven and overfitting may have occurred.[13] To correct for this overfitting, the regression coefficients of type of surgery should have been extra shrunk. We therefore did an additional analysis, in which we shrunk the regression coefficients of the four indicator variables for type of surgery with the heuristic shrinkage factor of van Houwelingen, which was estimated at 0.85.[19] This shrinkage factor was calculated by taking into account the difference in χ^2 of the model with 27 surgical procedures and the model with the five categories of surgery, and the difference

in degrees of freedom. However, when we applied the model with the extra shrunk regression coefficients of type of surgery, we found a similar performance as with the original prediction model (data not shown).

Third, it is often stated that the performance of models (developed or updated) should not be tested by split sample methods, but rather by bootstrapping or several (randomly selected) cross validations[13;21;22] as chance findings are more likely to occur with the former. However, we explicitly used the split-sample method based on time to cohere with the natural chronology in practice. It is possible however that the large drop in AUC in the test set of the updated models could indeed be due to chance or unfortunate split sampling.

Finally, we left out two other possible updating approaches. The first is to rebuild a complete new model starting from no prior knowledge and thus selecting the most important predictors based on the new data only. We explicitly did not apply this approach as there is ample evidence that this leads to overfitted prediction models with doubtful reproducibility in subsequent patient samples, certainly when the validation set is relatively small as often is the case.[3;20] The second approach is to retain the predictors of the original model and assess the added value of new predictors. Addition of new predictors may particularly increase the model's discriminative ability. For our example study, additional predictors of severe postoperative pain that are mentioned in the literature could be body mass index, type of anesthesia and duration of surgery. However, we previously found that these had no predictive value in the derivation set.[8;23] Besides that, duration of surgery was only assessed postoperatively in the updating and test set and the preoperatively expected duration was not available. We note that in other situations it may be worthwhile to study the effect of adding predictors to an existing model, which can easily be done with method 3.

To conclude, in agreement with previous studies, our example study also demonstrates with empirical data that prediction models can be updated with simple recalibration methods. These simple recalibration methods improved the calibration to a similar extent as the revision methods that made more extensive adjustments to the original model. Discrimination was not improved by any of the methods. More generally speaking, we stress that before any model is applied in practice, we strongly advise that any model (diagnostic or prognostic) is first validated and, if necessary, updated. A validated (and, if necessary, updated) model may cautiously be applied in new patients that are similar to the patients in the derivation and validation sets. However, when the user has reasons to believe that the model may perform differently in the new patients, data of the new patients should first be collected to test the performance before the model can be applied in daily clinical practice. Any model may perform slightly different in every new patient sample due to sampling variation. In that situation, the model does not need to be updated. However, two questions remain: when is the difference in model performance large and is updating needed, and when has a model been sufficiently updated? So far, this particular methodological area of prediction research has not been explored. Future research should address the question how many validation studies and what type of adjustments are needed before it is justified to implement a prediction model into clinical practice.

Acknowledgement

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZonMw 016.046.360).

References

1 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130(6):515-524.
2 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19(4):453-473.
3 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23(16):2567-2586.
4 Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med 1986; 105(4):586-591.
5 Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med 1986; 146(1):81-83.
6 Hosmer D.W., Lemeshow S. Applied logistic regression. New York: John Wiley and Sons, Inc., 1989.
7 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999; 99(16):2098-2104.
8 Janssen KJM, Kalkman CJ, Grobbee DE, Bonsel GJ, Moons KGM, Vergouwe Y. Validation of a clinical prediction rule for severe postoperative pain in new settings. Article submitted, available on request 2007.
9 Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea with vomiting and economic analysis. Anesthesiology 2001; 95(3):616-626.
10 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005; 58(5):475-483.
11 Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30(6):473-483.
12 Moerman N, van Dam FS, Muller MJ, Oosting H. The Amsterdam Preoperative Anxiety and Information Scale (APAIS). Anesth Analg 1996; 82(3):445-451.
13 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(4):361-387.
14 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12):1255-1264.
15 Little R.A. Regression with missing X's: a review. J Am Stat Assoc 1992; 87:1227-1237.
16 Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993; 13(1):49-58.
17 Cox DR. Two further applications of a model for binary regression. Biometrika 1958; 45:562-565.
18 Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3(2):143-152.
19 Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990; 9(11):1303-1325.
20 van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995; 14(18):1999-2008.

21 Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall, 1993.
22 Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54(8):774-781.
23 Kalkman CJ, Visser K, Moen J, Bonsel GJ, Grobbee DE, Moons KG. Preoperative prediction of severe postoperative pain. Pain 2003; 105(3):415-423.

Appendix

Regression coefficients of the updated models

	Method 0*	Method 1	Method 2	Method 3	Method 4	Method 5
Intercept	-0.42	-1.40	-1.19	-1.43	-1.61	-1.38
Ambulatory surgery	-0.70	-0.70	-0.53	-0.63	-0.88	-0.79
Female gender	-0.004	-0.004	-0.003	-0.004	0.29	0.16
Age	-0.009	-0.009	-0.007	-0.008	-0.003	-0.01
Preoperative pain score	0.11	0.11	0.08	0.10	0.12	0.10
Low pain surgery	0.50	0.50	0.38	0.45	0.59	0.70
Moderate pain surgery	0.92	0.92	0.70	0.83	0.57	0.94
High pain surgery	1.05	1.05	0.79	0.95	0.82	1.00
Highest pain surgery	1.72	1.72	1.30	1.55	1.41	1.62
Expected incision size ≥ 10 cm	0.39	0.39	0.29	0.35	0.50	0.73
APAIS anxiety score	0.05	0.05	0.04	0.05	0.01	0.03
APAIS need for information	-0.05	-0.05	-0.04	-0.05	-0.02	-0.03
Ambulatory surgery *female gender	0.67	0.67	0.50	0.60	0.56	0.49
Ambulatory surgery *low pain surgery	-0.10	-0.10	-0.08	-0.57	-0.44	-0.17
Ambulatory surgery *moderate pain surgery	-0.47	-0.47	-0.35	0.41	0.70	0.04
Ambulatory surgery *high pain surgery	-0.07	-0.07	-0.05	0.24	0.55	0.15
Ambulatory surgery *highest pain surgery	-1.51	-1.51	-1.14	0.85	0.82	-0.30

* Original prediction rule

Chapter 4

Application

Chapter 4.1

Application of clinical prediction rules in daily practice: Methods to handle missing predictor values

Abstract

Background Clinical prediction models combine patient characteristics and test results to predict the presence of a disease (diagnosis) or the occurrence of an event in the future (prognosis). A physician applying a prediction model for a patient with a missing predictor value needs to be advised how to handle missing predictor values. We present six strategies that handle missing values and compare the effects on the predictive accuracy of a prediction model that predicts the presence of deep venous thrombosis (DVT).

Method We developed and externally validated a prediction model consisting of seven predictors in respectively 1295 and 532 primary care patients. In an application set (259 patients) we mimicked three scenarios, in which an important predictor (D-dimer test), a weaker predictor (difference in calf circumference) and both predictors simultaneously were missing. We used six strategies that handle the missing values: imputation of the value zero, mean imputation, subgroup mean imputation, multiple imputation, applying a submodel (consisting of only the observed predictors) derived in the derivation set, and applying the submodel derived with 'one step sweep'. We compared the accuracy of the strategies in the application set, by assessing the discrimination (ability to distinguish between patients with the outcome and patients without the outcome, quantified with the area under the Receiver Operating Characteristic curve (ROC area)) and the calibration (agreement between the predicted probabilities and observed frequencies, expressed by a slope and intercept, ideally equal to respectively 1 and 0).

Results When the strong predictor was missing, or both predictors simultaneously, multiple imputation was best capable of improving the discrimination. When the weak predictor was missing, the discrimination was not reduced by any of the strategies. When the strong predictor was missing, only subgroup mean imputation led to a slope close to the reference slope (1.02 versus 1.06). When the weak predictor was missing, all slopes were (nearly) equal to the reference slope. When the two predictors were missing simultaneously, none of the strategies resulted in a slope close to the reference slope. When the strong predictor was missing, or both predictors simultaneously, only multiple imputation resulted in an intercept close to the reference intercept. When the weak predictor was missing, application of the model derived without the difference in calf circumference and multiple imputation resulted in an intercept close to the reference intercept.

Conclusion Multiple imputation was best capable of improving the predictive accuracy of the prediction model when one or more predictor values were missing.

Introduction

Setting a diagnosis or prognosis in clinical practice commonly requires the combination of patient characteristics and test results, to estimate the probability that a disease or outcome is present (diagnosis) or will occur (prognosis). It is, therefore, a multivariable process. To assist physicians in this process, for numerous situations so-called clinical prediction models (or prediction rules or risk score) have been developed. For example, for estimating the prognosis of newborns the Apgar score[1] has been developed, for the prediction of heart disease in the general population the Framingham risk score[2], and for intensive care patients the Acute Physiology and Chronic Health Evaluation (APACHE) score[3]. Roughly three consecutive phases can be distinguished in clinical prediction research; derivation of the prediction model, validation of the model in new subjects, and its application in individual patients in daily practice.[4-7]

When deriving or validating a prediction model in scientific studies, commonly more or less predictor values are missing. Usually, researchers conduct the so-called complete case analysis in which the data of the patients with missing values is simply neglected. However, complete case analysis leads to loss of power, but more importantly, the study results are often biased. [8-18] As an alternative, predictors with (many) missing values are often excluded from the analyses. This, however, may similarly lead to biased results. A more advanced and increasingly used method in epidemiological research is multiple imputation, a technique that uses all observed patient information to (multiply) impute the missing patient values, and subsequently allows for standard statistical analysis.[8-22] Multiple imputation has been advocated as a proper method to validly handle missing values in epidemiological – i.e. etiologic, prediction and therapeutic – research.[8-16;19-22]

Obviously, physicians who apply prediction models for their patients to assess a particular risk, may also face missing predictor values. For example, a risk score to predict the presence of a bacterial infection in children with acute fever without apparent source includes the predictor 'duration of fever'.[23] It may well be that the parents do not have measured the temperature or do not remember the exact duration of the fever. But, to apply a prediction model, all predictors in fact need to be known. A model can not be applied by simply leaving out that predictor, as the relative weights of the other predictors in the model may become invalid. Simply imputing the mean value of that predictor - e.g. obtained from their own population or from the population from which the model was developed - may also be insufficient and lead to invalid predictions and thus compromise patient care.[11;17;22] It is thus less straightforward how to handle missing predictor values when applying a prediction model in practice.

We present six strategies that can handle missing values when applying a prediction model in practice. In four strategies, the original prediction model is applied and the missing predictor values are imputed, and in the other two strategies, an adjusted prediction model is applied in which the predictor with missing values is excluded. We apply the strategies in empirical data for a prediction model that aimed to predict the presence or absence of Deep Venous Thrombosis (DVT).

Methods

The clinical example we used concerns the prediction of presence or absence of Deep Venous Thrombosis (DVT). Timely recognition or ruling out DVT in patients suspected of DVT is important because patients with untreated DVT may develop pulmonary embolism whereas unjustified therapy with anticoagulants poses a risk for major bleeding.[24] In primary care, the physician needs to decide on patient history, physical examination and usually the D-dimer assay result, which patients need to be referred to the hospital and which can be safely kept under their own surveillance. A diagnostic prediction model could aid physicians in this decision.

We used a cohort of 2086 patients suspected of DVT that has been described previously in the literature.[25;26]. The patients all underwent the same measurements and were selected from the same primary care practices. As said, prediction models are derived from derivation sets and subsequently tested in new patients in (usually smaller) validation sets.[4-7] To adhere to this process, we split the cohort according to previously used chronological subsets[25;26] into a derivation set, a validation set and a application set.

Derivation and validation of the prediction model

The derivation set consisted of 1295 primary care patients, included between January 2001 and May 2003, with a suspicion of DVT that participated in a diagnostic study that was previously presented (Table 1, second column).[25;27;28] After obtaining the patient history, the physical examination and the D-dimer result, all patients underwent repeated leg ultrasound as the reference method to determine the true presence or absence of DVT. The prediction model was developed with multivariable logistic regression and included seven predictors: age, absence of a leg trauma, vein distension, duration of symptoms, immobilisation, difference in calf circumference, and the D-dimer level. The prediction model was:

$$\log\left(\frac{\text{risk of DVT}}{1-\text{risk of DVT}}\right) = \text{linear predictor} =$$

-14.84 + 0.81*Absence of trauma-0.02*Age+0.39*Vein distension – 0.02*Duration of symptoms + 0.34*Immobilisation + 0.80*Log(difference in calf circumference) + 1.72*Log(D-dimer level) (Formula 1)

The risk of DVT in an individual patients (scale: 0-100%) can thus be calculated by:

$$\text{risk} = \frac{1}{1+e^{-\text{linear predictor}}} *100\%.$$

We then validated this prediction model in the second part of our data, i.e. 532 consecutive patients included between June 2003 and June 2005 (Table 1, third column).

Table 1 Patient characteristics in the derivation set, the validation set and the application set, n (%) unless stated otherwise.

Patient characteristics	Derivation set (n=1295)	Validation set (n=532)	Application set (n=259)
Age, years[a]	60 (18)	60 (17)	59 (18)
Male gender	465 (36%)	217 (41%)	86 (33%)
Oral contraceptive use	129 (10%)	46 (9%)	29 (11%)
Duration of symptoms, days[a]	8 (9)	7 (6)	9 (12)
Leg trauma absent	1104 (85%)	438 (82%)	213 (82%)
Malignancy present	77 (6%)	19 (4%)	18 (7%)
Immobilisation	172 (13%)	65 (12%)	41 (16%)
Recent surgery	163 (13%)	66 (12%)	31 (12%)
Swelling whole leg	580 (45%)	231 (43%)	121 (47%)
Vein distension	233 (18%)	89 (17%)	48 (19%)
Log(Calf circumference)[a]	1.14 (0.59)	1.13 (0.58)	1.16 (0.57)
Log(D-dimer level)[a]	6.80 (1.21)	6.93 (1.15)	6.66 (1.28)
DVT present	289 (22%)	91 (17%)	35 (14%)

[a] Mean (standard deviation)

Application of the DVT model

The application set included the last included 259 patients suspected of DVT (Table 1, fourth column). This data set originally did not contain missing predictor values, and served as the reference situation in which we could apply the original prediction model to all patients. We now mimicked three scenarios in which predictor values were missing for the individual patients. In the first scenario, the D-dimer value (strongest predictor) was missing for all patients. In the second scenario, the difference in calf circumference (weaker predictor) was missing for all patients. In the third scenario, both predictors were missing for all patients. We used the D-dimer value and the difference in calf circumference for this purpose, as we figured that in practice it may well occur that physicians would not measure these parameters due to lack of resources or time.

Strategies to handle missing values when applying a prediction model

We compared six strategies (Table 2) that can handle the continuous missing predictor values when applying an existing prediction model in daily practice; four strategies impute the missing value and two strategies apply an adjusted prediction model; i.e. the submodel without the predictor with missing values.

1. Imputation of the value zero

The missing predictor value was imputed with the value 'zero'. This in fact means - for example in the first scenario - that the predictor D-dimer value is simply excluded from the model and the intercept and regression coefficients of the remaining predictors in Formula 1 are used without adjustments.

Table 2 Strategies to handle missing predictor values when a prediction model is applied in daily clinical practice.

Missing predictor values imputed	
1. Zero imputation	The missing predictor value is imputed with zero, implying that the predictor is ignored and the prediction model is applied with the unadjusted regression coefficients of the original model.
2. Mean imputation	The missing predictor value is imputed with the mean value, estimated in the derivation set.
3. Subgroup mean imputation	The missing predictor value is imputed with a subgroup mean value, as estimated in the derivation set. The subgroups were determined by gender and five age categories.
4. Out of sample imputation	The missing predictor value is imputed with multiple imputation techniques. Each patient record with missing predictor values was individually merged to the derivation set to apply multiple imputation.
Prediction model adjusted	
5. Submodel without predictors with missing values, derived in the derivation set	The submodel contains adjusted regression coefficients that are estimated in the derivation set.
6. Submodel without predictor(s) with missing values, derived with one step sweep	The submodel contains adjusted regression coefficients that are estimated with the original regression coefficients and covariance matrix.

2. Mean imputation

The missing predictor value was imputed with the mean value of the predictor, as estimated in the derivation set. For example, when the D-dimer value was missing (first scenario), the mean log(D-dimer value) of the patients in the derivation set was imputed.

3. Subgroup mean imputation

The missing predictor value was imputed with a subgroup mean value, estimated in the derivation set. Subgroups were determined by gender and five age categories. For example, when the D-dimer value in the application set was missing (first scenario) for a male patient of 44 years old, the mean log(D-dimer value) of male patients between 40 and 50 years of age in the derivation set was imputed.

4. Multiple imputation

The missing predictor value was imputed with multiple imputation techniques using *mice* [29]. Multiple imputation techniques estimate multiple values of the missing predictor, based on all other available predictors and patient characteristics of the patients.[8-22] This method is rather straightforward and feasible when analysing a dataset. But to use this method in practice when a prediction model is actually applied in an individual patient, access to the individual data of the derivation set is required. Hence, the derivation set needs to be available to the users of the prediction model (for example via a website). Accordingly, the record of the individual patient with missing predictor values is in fact added to the

derivation set, and multiple imputation can be applied. In our example, we imputed ten values of the missing predictor in each patient of the application set. Then for each patient ten linear predictors were calculated with Formula 1, which were then averaged to obtain the risk of the presence of DVT.

5. Submodels derived from the derivation set

We derived submodels without the predictor(s) with missing values in the derivation set. Hence, the intercept and regression coefficients of the observed predictors of the original prediction model are adjusted for the exclusion of the predictor with missing values. For a patient with missing data in predictors x_2, we thus formulated a submodel consisting of only the observed predictors x_1:

$$\log\left(\frac{\text{risk of DVT}}{1-\text{risk of DVT}}\right) = \beta_1 x_1 + \beta_2 x_2$$

were β_1 is the vector of the intercept and regression coefficients of the observed predictors, x_1 are the observed predictor values, $\beta_2 = 0$ and x_2 are the unobserved predictor values. For each scenario, the risk of DVT of the individual patients of the application set was estimated by the corresponding submodel.

6. Submodels derived by the One step sweep method

As in strategy 5, submodels were derived that contain only the predictors with observed values, but here with a non-iterative one-step approximation, called the 'one step sweep' method.[30] The adjusted regression coefficients of the submodel were based on the regression coefficients of the original model (see Formula 1) and the covariance matrix as estimated in the derivation set. We will briefly explain this method, and refer to the literature for a more profound description.[30] The one step sweep strategy separates the vector of regression coefficients into

$$\begin{pmatrix}\beta_{\text{submodel}} \\ \beta_{\text{unobserved predictor(s)}}\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}, \begin{bmatrix}V_{11} & V_{12} \\ V_{21} & V_{22}\end{bmatrix}\right)$$

in which the β's are the intercept and regression coefficients of the predictors, and V is the estimated covariance matrix of the original prediction model. V_{11} refers to the covariances between the observed predictors, V_{12} and V_{21} to the covariances between the observed and the unobserved predictors (V_{12} and V_{21} contain the same elements, $V_{12} = V_{21}{}^{T}$), and V_{22} to the covariances between the unobserved predictors.

When only one predictor contains with missing values (our scenario 1 and 2), V_{22} simply refers to the variance of the regression coefficient of that missing predictor. When more predictors contain missing values (scenario 3, both D-dimer and difference in calf circumference missing), V_{22} refers to the covariance matrix of the regression coefficients of these predictors with missing values. Hence, according to the one-step-sweep method, the intercept and regression coefficients of the submodel that includes only predictors with observed values (x_1) can be estimated as

$\beta_{\text{submodel}} = \beta_1 - V_{12} * V_{22}{}^{-1} * \beta_2$

were β_1 is the vector of the original intercept and regression coefficients of the observed

predictors, and $V_{12}*V_{22}*\beta_2$ a vector of 'correction factors' based on the covariances between the observed and the unobserved predictors as estimated in the derivation set (V_{12}), the original covariances between the unobserved predictors (V_{22}), and the original regression coefficient(s) of the unobserved predictors (β_2).

Accuracy measures

We estimated the accuracy by assessing the discrimination and calibration. Discrimination is the ability of the model to distinguish between patients with the outcome and patients without the outcome. It can be quantified with the area under the Receiver Operating Characteristic curve (ROC area), which is equal to the c-statistic for a dichotomous outcome variable.[31] An ROC area ranges from 0.5 (no discrimination; same as flip of a coin) to 1.0 (perfect discrimination).[32]

Calibration refers to the agreement between the predicted probabilities and observed frequencies of the outcome. It can be graphically assessed with a calibration plot with the predicted probabilities on the x-axis and the observed frequencies on the y-axis. The calibration plot shows a calibration line, which can be described with a calibration slope and a calibration intercept.[33;34] These are estimated by fitting the linear predictor of the applied model (original or adjusted) as the only covariate in a logistic regression model. The calibration slope and calibration intercept of a model in new patients are ideally equal to 1 and 0 respectively (perfect calibration). A slope < 1 indicates optimism (predictions are too extreme), while a slope > 1 indicates that the predictions are not extreme enough. When the slope is not equal to 1, the interpretation of the intercept is difficult. Hence, we estimated the intercept with the slope fixed at 1 (calibration in the large). When this intercept is close to 0, the calibration in the large is good, i.e. the mean predicted probability equals the mean observed frequency.

Results

Prerequisites for applying the six strategies

Strategy 1 was the only strategy that we could apply without additional estimations of the derivation of submodels. The prerequisites for the other strategies will be presented.

Strategy 2. The overall mean value of log(D-dimer) level and the log(difference of calf circumference) in the derivation set were 6.83 and 1.14 respectively.

Strategy 3. The subgroup means, for log(D-dimer level) and log(difference of calf circumference) in the derivation set are presented in the appendix, Table 1. For both the D-dimer level and the difference in calf circumference, the subgroup means differed from the mean of the whole population and were higher for male and older patients.

Strategy 4. We stored the derivation set and multiple imputation script for the multiple imputation strategy (script available on request).

Strategy 5. Three submodels were derived in the derivation set in which respectively the D-dimer, the difference in calf circumference, and the two predictors simultaneously were excluded, Table 3 (third, fifth and seventh column).

Strategy 6. Three submodels were derived with the one step sweep strategy, Table 3 (fourth, sixth and eighth column). Table 2 of the appendix shows the covariance matrix of the regression coefficients of the original model that was used in the one step sweep strategy.

Table 3 Intercept and regression coefficients of the predictors of the original prediction model (applied in strategy 1-4), and the submodels without the predictor(s) with missing values, derived in the derivation set (strategy 5) or with one step sweep (strategy 6).

Missing predictor(s)	None	D-dimer level		Calf circumference		D-dimer level and calf circumference	
	Original model	Derivation set	One step sweep	Derivation set	One step sweep	Derivation set	One step sweep
Strategy	**1-4**	**5**	**6**	**5**	**6**	**5**	**6**
Predictors							
Intercept	-14.84	-2.89	-2.66	-13.83	-13.50	-2.12	-1.62
Trauma	0.81	0.41	0.54	0.64	0.63	0.28	0.37
Age	-0.02	0.001	-0.005	-0.013	-0.02	0.005	-0.001
Vein	0.39	0.39	0.46	0.38	0.38	0.42	0.46
Duration	-0.02	-0.009	-0.01	-0.017	-0.02	-0.01	-0.01
Immobilisation	0.34	0.02	-0.07	0.33	0.33	0.04	-0.08
Calf	0.80	0.73	0.85	-	-	-	-
D-dimer	1.72	-	-	1.68	1.68	-	-

Trauma Absence of trauma
Vein Vein distention
Duration Duration of symptoms
Calf Log(calf circumference)
D-dimer Log(D-dimer level)

Accuracy of the six strategies in the application set

Discrimination

When missing predictor values were not missing in the application set (reference situation), the ROC area of the original DVT model was 0.90 (95% CI: 0.84-0.96) (Table 4). When the D-dimer level was missing (first scenario), the ROC area decreased to approximately 0.70 in all strategies, except with multiple imputation (ROC area = 0.77). When the difference in calf circumference was missing (second scenario), the ROC did not decrease in any of the strategies (ROC area = 0.89 or 0.90). When both the D-dimer level and the difference in calf circumference were missing (third scenario), the ROC area decreased to an ROC area below 0.65 for all strategies, except with multiple imputation (ROC area = 0.78).

Calibration

When no values were missing in the application set, the calibration slope was 1.06 (reference situation). When the D-dimer level was missing (first scenario), only subgroup mean imputation resulted in a calibration slope (1.02) close to the reference slope (Table 5). When the difference in calf circumference was missing (second scenario), all slopes were similar to the reference slope. When the two predictors were missing simultaneously (third scenario), none of the strategies led to calibration slopes close to the reference slope, though subgroup imputation resulted in the smallest deviation (slope = 0.94).

When no values were missing in the application set (reference situation), the intercept given that the slope was equal to 1 was -0.10. When the D-dimer level was missing (first scenario), all strategies led to insufficient calibration in the large (intercept not equal to 0), apart from multiple imputation (intercept = -0.06). When the difference in calf

Table 4 Effect of the six strategies on the discriminative ability of the prediction model in the application set, expressed by the ROC area (95% confidence intervals) when the D-dimer value is missing (scenario 1), when a difference in calf circumference is missing (scenario 2), and when the two predictors are simultaneously missing (scenario 3). The ROC area when no data were missing in the application set (reference situation) was 0.90 (95% confidence interval 0.84 – 0.96).

	Predictor with missing data		
	D-dimer	**Difference in calf circumference**	**D-dimer and difference in calf circumference**
1. Zero imputation	0.70 (0.61-0.79)	0.89 (0.83-0.96)	0.62 (0.53-0.71)
2. Mean imputation	0.70 (0.61-0.79)	0.89 (0.83-0.96)	0.62 (0.53-0.71)
3. Subgroup mean imputation	0.69 (0.61-0.78)	0.90 (0.83-0.96)	0.64 (0.55-0.74)
4. Multiple imputation	0.77 (0.69-0.84)	0.90 (0.84-0.96)	0.78 (0.71-0.86)
5. Model without predictor(s) with missing values, estimated in the derivation set	0.70 (0.62-0.79)	0.89 (0.83-0.96)	0.64 (0.54-0.74)
6. Model without predictor(s) with missing values, estimated by One step sweep	0.70 (0.61-0.78)	0.89 (0.83-0.96)	0.66 (0.57-0.75)

Table 5 Effect of the six strategies on the calibration of the prediction model in the application set, expressed by the slope and the intercept when the calibration slope was fixed at 1, either when D-dimer values are missing (scenario 1), a difference in calf circumference (scenario 2), or the two predictors are simultaneously missing (scenario 3). The slope and intercept with the slope fixed at 1 when no data were missing in the application set (reference situation) were 1.06 and -0.10 respectively.

		Predictor with missing data		
		D-dimer	**Difference in calf circumference**	**D-dimer and difference in calf circumference**
1. Zero imputation	Slope	1.13	1.05	0.89
	Intercept[a]	12.48	0.97	13.46
2. Mean imputation	Slope	1.13	1.05	0.89
	Intercept[a]	0.73	0.05	0.80
3. Subgroup mean imputation	Slope	1.02	1.06	0.94
	Intercept[a]	0.72	0.06	0.82
4. Multiple imputation	Slope	0.76	1.07	0.83
	Intercept[a]	-0.06	-0.04	0.01
5. Model without predictor(s) with missing values, estimated in the derivation set	Slope	1.47	1.07	2.14
	Intercept[a]	-0.30	-0.03	-0.28
6. Model without predictor(s) with missing values, estimated by One step sweep	Slope	1.27	1.04	2.11
	Intercept[a]	-0.42	0.11	-0.42

[a] Intercept when the calibration slope was fixed at 1

circumference was missing (second scenario), calibration in the large was similar to the reference situation when the submodel was applied that was derived in the derivation set without the difference in calf circumference (intercept = -0.03) and when multiple imputation was applied (intercept = -0.04). Zero imputation led to the worst calibration in the large (intercept = 0.89). When the two predictors were missing simultaneously (third scenario), all strategies resulted in insufficient calibration in the large, apart from multiple imputation (intercept = 0.01).

Discussion

We presented six strategies to handle missing values when a prediction model that predicts the presence of DVT is applied in daily clinical practice. We compare the effects of these six strategies on the predictive accuracy of the prediction model when a strong predictor, a weaker predictor, or both predictors simultaneously were missing in daily clinical practice. Multiple imputation showed the best accuracy in our study.

In scenario 1 and 3 (strong predictor missing, and both predictors missing simultaneously), multiple imputation resulted in the highest ROC area, though lower than the reference situation. Imputation of the value zero led to the lowest ROC area, which was expected as a predictor is excluded from the prediction model without adjustment of the intercept and the regression coefficients of the other predictors. Imputation of the mean resulted in a similar ROC area, as this strategy results in the same rank order of the predicted probabilities as the zero imputation. Imputation of the subgroup mean can hypothetically improve the discrimination, although this was only shown for scenario 3. Further, the submodels that contain only the predictors with observed values (either derived in the derivation set or with the one step sweep strategy) can hypothetically better discriminate than the simple imputation strategies 1 and 2 (respectively zero and mean imputation). Again, this was only shown in scenario 3. Note that when a strong predictor is missing, as for example the D-dimer test in our study, a submodel without this predictor will hardly ever reach the same discriminative ability as the full model. When the relatively weak predictor was missing (scenario 2), all strategies led to ROC areas similar to the reference situation. Apparently, the discriminative ability of the model is mostly based on the strong predictor D-dimer. The reason that the model also includes other predictors is that the selection of predictors in the derivation of the model was based on the likelihood of the model. [35;36] This is a measure of model fit, in which both the discrimination and calibration of the model is reflected. If the selection of predictors would have been solely based on the discriminative ability of the model, less predictors (if any) would have been included in the model additional to the D-dimer.

We expected that the slope would be larger than 1 (predictions not extreme enough; i.e. low predictions are too high and high prediction too low) after no imputation and mean imputation, and would improve after subgroup mean imputation. Our rationale was that the regression coefficients of the observed predictors of the model would have been larger if the model would have been derived without the missing predictor (given that there is correlation between the predictor with missing values and the other predictors). Imputation of the subgroup mean can improve the slope, compared to zero and mean imputation. For example, consider a risk increasing predictor with missing values. Patients with low predicted probabilities (based on the observed predictors) may be in a subgroup with low imputed values, patients with high predicted probabilities may be in a subgroup

with high imputed values. This would result in more extreme predictions in patients. Last, we expected that the submodels of predictors with observed values (either derived in the derivation set, or derived by the one step sweep strategy) and multiple imputation would result in a slope close to the reference slope. When the strong predictor was missing, or both predictors simultaneously (scenario 1 and 3), imputation of the subgroup mean resulted in a slope closest to the reference slope (1.06). In contrast to our expectations, we found a slope > 1 for the submodels (either derived in the derivation set, or by the one step sweep strategy), and a slope < 1 for multiple imputation. When the difference in calf circumference was missing (second scenario), all slopes were (nearly) equal to the reference slope (1.06).

When no values were missing in the application set (reference situation), the calibration in the large (intercept given that the slope is equal to 1) was -0.10. As expected, zero imputation resulted in the worst calibration in the large, as excluding a predictor from a model without adjustment of the intercept and regression coefficients of the other predictors leads to systematically overestimated or underestimated probabilities. Imputation of the mean and the subgroup mean resulted in calibration in the large closer to the reference situation, as the effect of the missing predictor is incorporated in the risk estimation. Application of the submodels, either derived in the derivation set or by the one step sweep strategy, resulted in calibration in the large closer to the reference situation than (subgroup) mean imputation. Also as expected, multiple imputation led to the best calibration in the large, as this strategy best approaches the missing predictor values.

We did not expect some of our findings. First, imputation of the subgroup mean did not always result in an ROC area closer to the reference situation than imputation of the mean. This suggests that the variation in the imputed values of the subgroup means compared to the single imputed value of the mean imputation was not large enough. Subgroups that are based on other predictors may increase the variation in the imputed values and improve the discrimination. Second, the submodels without the predictor with missing values (either derived in the derivation set or by the one step sweep strategy) insufficiently improved the calibration in the large and led to the worst slopes, compared to the other strategies. The effect of these strategies on the accuracy measures will probably depend on the data and the models, and may be different in other situations. Third, multiple imputation did not always lead to a good slope. Future research should examine why this can occur. Shrinkage of the imputation models may be a possible solution.

When interpreting the results of our study, some considerations need to be taken into account. First, our results are based on one empirical example. Other datasets with other prediction models predicting other outcomes may show different results. For example, the submodels (derived in the derivation set or by the one step sweep strategy) may lead to better results when the predictors of the model have a similar predictive strength. In our study, the D-dimer was such a strong predictor that estimating a submodel without this model would always lead to reduced discrimination. Second, we did not show all possible strategies to handle missing values when applying a prediction model in daily clinical practice. For example, we could have fitted regression models to estimate the missing values, in which the predictor with missing values is the dependent variable and the other predictors the independent variables. However, note that for a prediction model with seven predictors like ours, one would need to develop and store the $6*2^7=768$ potential regression models. Another strategy would be to derive and store all possible

submodels without the predictor(s) with missing values. In our study, we considered two predictors that could have missing values. However, all seven predictors hypothetically can have missing values. This means that in our example, $2^7 = 128$ submodels would have to be developed and stored, at substantial computational cost. The one step sweep strategy can estimate these submodels without the need to develop and store all the submodels.[30] Third, mean imputation and subgroup mean imputation are easily applicable in daily clinical practice. Also, when the submodels derived in the derivation set and the covariance matrix necessary for the one step sweep are presented in the literature, these submodels can be easily applied. Only for multiple imputation, a personal computer is a necessity. As a result, physicians might be reluctant to apply this strategy. However, with the evolving introduction of the electronic patient records in primary and secondary care with its potential for built in algorithms, these strategies may be more easily implemented and applied. Fourth, the gain of multiple imputation over single imputation with regression techniques is in the correct estimation of the standard errors, in our example the standard errors of the predicted probabilities. We did not take full profit of this advantage, as in our study the interest was not in the confidence intervals of the predicted probabilities but in the predictive accuracy of the model. However, in situations where the confidence intervals of predicted probabilities are of interest, for example when the estimation of the confidence intervals becomes mainstream practice in daily clinical care, this will be an extra advantage of the application of multiple (multiple) imputation. Fifth, we could have split the cohort of patients into a derivation set and an application set, which would have resulted in a larger derivation set. However, in our study we explicitly wanted to use a validation set to test the accuracy of the newly developed model. Although this is the recommended strategy, it is still rarely applied in daily practice. We would like to stress that before any model is applied in daily clinical practice, the model needs to be tested in new patients.[6;7]

In conclusion, when a prediction model is applied in daily clinical practice and a predictor has missing values, excluding that predictor without adjustments of the intercept and the regression coefficients of the observed predictors should be avoided as the discrimination and calibration can be dramatically reduced. Imputation of the mean can not improve the discrimination, although the calibration in the large can be improved. Imputation of the subgroup mean may improve the discrimination and the calibration in the large, if the variance between the values in the subgroups is large enough. Estimating submodels without the predictor, either by estimation of the regression coefficients without the predictor with missing values in the derivation set, or by the one step sweep strategy might improve the predictive accuracy, but may as well worsen it. Multiple imputation was in our study best capable of improving the discrimination and the calibration in the large.

Acknowledgement

Part of this work has been conducted in the Department of Statistics, Harvard University (Professor DB Rubin) and in the Department of Biostatistics, Vanderbilt University Medical School (Professor FE Jr Harrell). We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZonMw 016.046.360).

References

1 AULD PA, Rudolp AJ, AVERY ME, CHERRY RB, DRORBAUGH JE, KAY JL et al. Responsiveness and resuscitation of the newborn. The use of the Apgar score. Am J Dis Child 1961; 101:713-24.:713-724.

2 Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976; 38(1):46-51.

3 Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991; 100(6):1619-1636.

4 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19(4):453-473.

5 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(4):361-387.

6 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130(6):515-524.

7 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006; 144(3):201-209.

8 Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003; 56(1):28-37.

9 Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995; 48(2):209-219.

10 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12):1255-1264.

11 Little R.J. Regression with missing X's: a review. J Am Stat Assoc 1992; 87:1227-1237.

12 Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, New Jersey: John Wiley & Sons; 1987.

13 Little RJ. Methods for handling missing values in clinical trials. J Rheumatol 1999; 26(8):1654-1656.

14 Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, New Jersey: John Wiley & Sons; 1987.

15 Rubin DB. Multiple Imputation after 18+ years. Journal of the American Stat Association 1996; 91:473-489.

16 Schafer JL. Analysis of Incomplete Multivariate Data. Chapman & Hall /CRC; 1997.

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

18 Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10):1092-1101.

19 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10(4):585-598.

20 Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research 1998; 33:545-571.

21 Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999; 8(1):3-15.

22 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7(2):147-177.

23 Bleeker SE, Moons KG, Derksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. Acta Paediatr 2001; 90(11):1226-

1232.
24 Hirsh J, Hoak J. Management of deep vein thrombosis and pulmonary embolism. A statement for healthcare professionals. Council on Thrombosis (in consultation with the Council on Cardiovascular Radiology), American Heart Association. Circulation 1996; 93(12):2212-2245.
25 Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005; 94(1):200-205.
26 Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006; 55(7):613-618.
27 Oudega R, Moons KG, Hoes AW. Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care. Fam Pract 2005; 22(1):86-91.
28 Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med 2005; 143(2):100-107.
29 van Buuren S, Oudshoorn C. http://web inter nl net/users/S van Buuren/mi/hmtl/mice htm) [2007 Available from: URL:http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm).
30 Marshall G, Warner B, MaWhinney S, Hammermeister K. Prospective prediction in the presence of missing data. Stat Med 2002; 21(4):561-570.
31 Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3(2):143-152.
32 Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148(3):839-843.
33 Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993; 13(1):49-58.
34 Cox DR. Two further applications of a model for binary regression. Biometrika 1958; 45:562-565.
35 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007; 115(7):928-935.
36 Harrell FE, Jr. Regression modelling strategies. Springer-Verlag, New York; 2001.

Appendix

Table 1 Mean log(D-dimer level) and mean log (difference in calf circumference) for gender and age specific subgroups estimated in the derivation set. These values are used in strategy 3 (subgroup mean imputation).

Log(D-dimer) level	Female mean	Male mean
Younger than 45 years	6.33	6.54
Between 45 and 54 years	6.48	6.66
Between 55 and 64 years	6.58	6.99
Between 65 and 74 years	6.69	7.32
75 years or older	7.19	7.38
Log(Difference in calf circumference)		
Younger than 45 years	0.94	1.07
Between 45 and 54 years	1.07	1.21
Between 55 and 64 years	1.12	1.16
Between 65 and 74 years	1.13	1.23
75 years or older	1.25	1.25

Table 2 Covariance matrix of the regression coefficients estimated in the derivation set

	Intercept	Trauma	Vein	Age	Duration	Immobilisation	Calf	D-dimer
Intercept	0.902	-0.083	0.000	-0.006	-0.001	-0.049	-0.029	-0.098
Trauma	-0.083	0.072	-0.002	-0.001	0.000	0.017	0.005	0.002
Vein	0.000	-0.002	0.044	-0.001	0.001	0.002	0.000	-0.001
Age	-0.006	-0.001	-0.001	0.003	0.000	0.000	-0.001	-0.001
Duration	-0.001	0.000	0.001	0.000	0.003	0.001	0.000	0.000
Immobilisation	-0.049	0.017	0.002	0.000	0.001	0.073	0.000	0.003
Calf	-0.029	0.005	0.000	-0.001	0.000	0.000	0.027	-0.001
D-dimer	-0.098	0.002	-0.001	-0.001	0.000	0.003	-0.001	0.014

Trauma	Absence of a leg trauma
Vein	Vein distention
Age	Age per 10 years
Duration	Duration of symptoms per 5 days
Calf	Log(difference of a calf circumference)
D-dimer	Log(D-dimer level)

Chapter 5

Discussion

Electronic patient records: patient care as a basis for clinical prediction research and vice versa

Electronic patient records (EPR) are increasingly used in medical care to replace conventional paper files by medical records in a digital format. The EPR facilitates storages and retrieval of data on medical history, diagnosis, treatment and prognosis, provides secure access of patient information by clinical staff at any given location, and can be used to standardise care pathways, guidelines and protocols. While the primary aim of the EPR is to aid patient care it creates highly attractive opportunities for research, in particular diagnostic and prognostic prediction studies. We will discuss how routine medical data stored in the EPR can be used for research to develop and validate clinical diagnostic and prognostic prediction models, and how in turn the EPR may promote the implementation and updating of such models in routine care.

Clinical prediction research

Diagnostic and prognostic probability estimations by physicians are often made implicitly, based on clinical knowledge and experience. Carefully developed and validated prediction models are meant to guide these estimations more explicitly. In clinical prediction research, patient characteristics (including symptoms, signs, patient demographics, disease characteristics, test results, and received interventions) are combined in statistical models to estimate the probability that a disease or outcome is present (diagnosis) or will occur (prognosis).[1;2] The predicted probabilities of these prediction models can be used for several purposes (Table 1).

Table 1 Objectives, motives and examples of the application of clinical prediction models

Objective	Motive	Example
Informing patients and their families	To discuss the probabilities of the presence of a disease or the occurrence of an event in the future	Applying the Framingham risk score to discuss the 10 year risk of cardiovascular disease[3]
Assisting in medical decision making	To assist individual patients in intervention choices	Balancing the risks of mortality after elective surgery (predicted by a prediction model) and after rupture in patients with abdominal aortic aneurysms[4]
Stratification of patients	To stratify patients by disease severity to create risk groups for randomized controlled trials	Testing the efficacy of tamoxifen in breast cancer patients with a poor prognosis, selected by a prediction model[5]
Correcting for different prognostic profiles	To assess the quality of care of physicians, clinical departments or hospitals by comparing the observed outcome frequencies, corrected for different prognoses	Comparing the mortality in neonatal intensive care units, corrected for prognosis, with the predicted mortality by the CRIB score[6] (clinical risk index for babies)

Several consecutive phases can be distinguished in clinical prediction research; development of the prediction model, validation of the model in new subjects (and updating if necessary), impact analyses, and the application in daily clinical practice. In the development phase, the predictive strength of potential predictors is assessed. The predictive per-

formance of most developed prediction models is decreased when tested in new patients, compared with the performance estimated in the patients that were used to develop the model.[3-5] Therefore, before a prediction model can be applied in daily clinical practice, it needs to be tested in new patients (i.e. externally validated). When the predictive performance is decreased in new patients, the model can be updated by combining the information of the model with data of the new patients. As a result, the updated models are adjusted to the particular characteristics of the new patients. Finally, before models can be applied in daily clinical practice with confidence, the impact of the validated model to change physicians decisions and improve patient outcome may need to be estimated.[5]

The EPR in practice

For physicians, the primary objective to work with an EPR is to aid patient care and to increase the efficiency of their care. Information that is relevant for their patient management - including diagnosis, prognostic and treatment - is collected and registered. As a result, and in contrast to data collection in typical research settings (such as randomised therapeutic trials), the data collected in the EPR realistically reflect the consecutive steps in patient management, in agreement with daily clinical practice. Moreover, the quality and completeness of the data reflect the quality of the data collected in routine patient care, and thus the quality and completeness that can be achieved when for example prediction models are implemented in the EPR for use in individual patients.

In clinical practice, a patient visits a physician with a particular set of symptoms and signs. Results from the diagnostic and prognostic work-up, including the final diagnosis and (if applicable) initiated treatment are documented. In general, there is variation among physicians in the level of detail with regard to patient characteristics that is recorded.

The patient characteristics are ideally registered according to standard (international) classifications. For example, in primary care, symptoms, signs and diagnoses are ideally registered according to the International Classification of Primary Care (ICPC). For secondary care, there is no such classification for the symptoms and signs, but for diagnoses the International Classification of Disease (ICD-10) is generally available. The use of standard international classifications should be promoted to improve standardization in electronic registration of routine care and to enhance medical research.

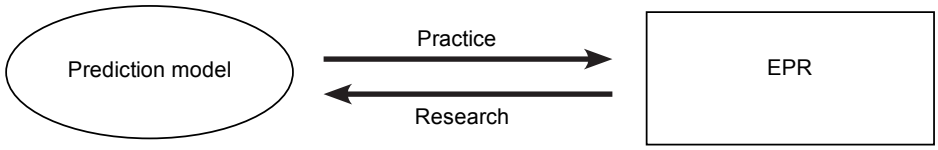
The EPR in research: development and validation of prediction models

Diagnostic research starts with a sample of patients presenting with particular symptoms or signs that rise the suspicion of a particular disease. When a diagnostic model is developed, patients are selected based on the symptoms and signs or suspected disease, and not on the final diagnosis.[6-8] Prognostic research starts with a sample of patients (usually with a particular disease or diagnosis) at risk of a certain event in the future. Hence, to properly select the patient groups for diagnostic and prognostic research, the registration of the main reason for visit or referral (in case of secondary care) in the EPR is as important as the registration of the final diagnosis.

In the development phase of either a diagnostic or prognostic model, the predictive strength of potential predictors is assessed (Figure 1, Research arrow). Patient characteristics that are recorded in the EPR can be studied in relation to a diagnostic or prognostic outcome. Therefore, the EPR should comprise all patient characteristics, preferably in a

consistent manner. Consider for example a diagnostic model that predicts the presence or absence of deep venous thrombosis (DVT) in primary care patients with a suspicion of DVT.[9] When the DVT model was developed, 17 potential predictors were considered. However, the final prediction model included 9 predictors: gender, duration of symptoms, absence of a leg trauma, presence of malignancy, immobilisation, pain when walking, oedema, difference in calf circumference and vein distension. This model could have been developed with EPR data, given that all potential predictor values and the final diagnosis were consistently documented.

Figure 1



Unfortunately, EPR data may comprise less information than would be ideal from a research perspective. Information that is needed for (prediction) research and not so much for patient care will not automatically be collected by physicians. Prediction models can be developed on data that have been collected prospectively and retrospectively. When data are obtained prospectively, the predictors and the outcome need to be documented from the patients before the model can be developed. Physicians can be instructed what the potential predictors of interest are to collect the relevant data. However, when the data are obtained retrospectively (the data have already been collected), researchers can only study the data at hand and often face many missing values.[10] Missing patient information (predictors and outcome) can severely distort scientific inferences from a dataset. However, there are opportunities to solve the problem of missing data. Multiple imputation is a statistical technique that uses all other observed patient information to fill in logical values for the missing data.[11-17] For example, patient characteristics that are not included in the prediction model may still be used to impute other predictor values that are missing. To achieve this, more patient information may be needed than the predictors eventually included in the model. Fortunately, this is typically the case in the EPR.

In the validation phase, the predictive performance of the prediction model is assessed in new patients. Again, prospective validation will in general be more feasible than retrospective validation, as in the former physicians can be properly instructed on the predictors (and outcome) of interest to collect the relevant data

The EPR in practice and research: impact analysis, implementation and updating of prediction models

Not only are EPR data potentially valuable to develop and validate clinical prediction models, implementation of prediction models in an EPR can also promote Evidence Based Medicine (Figure 1, Practice arrow). However, before models can be implemented and applied in daily clinical practice, the impact of the validated model to change physicians decisions and improve patient outcome may need to be estimated.[5]

Impact analysis

To determine whether a validated prediction model will actually be used by physicians, will change or direct physicians' decisions, and will improve patient outcomes or reduce costs, an impact study or impact analysis should be performed (Figure 1, Research arrow) [5;18]. In the ideal design of an impact study, physicians or care units are randomized to either the index group – which is 'exposed' to the use of the prediction model – or to the control group which uses 'care or clinical judgment as usual'.[18] Patients are followed up to determine the impact of the prediction model on patient outcome and on cost-effectiveness. The EPR can be a valuable tool in such studies, as the prediction model can be incorporated in the EPR for the physicians randomized to the index group, showing the predicted probabilities to these physicians only (Figure 1, Practice arrow).

Implementation

The prediction model (diagnostic or prognostic) can be incorporated in the EPR in such a way that when the values of all predictors are registered, the probability of having (diagnosis) or getting (prognosis) the outcome is automatically predicted and presented (Figure 1, Practice). Prediction models may be incorporated in the EPR in three ways. First, a model can be activated as soon as one of its predictors is registered. For example, when the symptom 'oedema in leg' is recorded in the EPR, the diagnostic model for estimating the probability of DVT presence or absence (discussed above) directly appears on the screen. The physician immediately sees which other predictors of the model need to be documented from the patient. Second, a model can be activated when a diagnosis (or the suspicion of a diagnosis) is registered. For example, when a physician registers DVT as a potential diagnosis in a patient with a swollen red leg, the diagnostic DVT model appears on the screen. Third, the model is actively called by the physician.

When incorporating a prediction model in the EPR, the probability prediction can be followed by a so called therapy advice. In general, a prediction model is called indicative when it presents the predicted probability, and it is called directive when the predicted probability is linked to a therapy advice. [5] For example, for the implemented primary care DVT model discussed above, a threshold can be introduced to refer a patient for additional testing.[19]A potential drawback of linking a therapy advice to a predicted probability is that it may be perceived as paternalistic or 'cookbook medicine'.[5] Therefore, before a model is incorporated in the EPR, the intended users should be consulted, so that it matches their needs. This can well be incorporated in an impact study.[5]

When a model has frequently been proven to be accurate in diverse populations, the more likely it is that the prediction model can be successfully applied in practice.[4;5;20] Yet, there are still reasons why the model is not as successful in daily practice. Physicians may *feel* that their often implicit estimation of a particular predicted probability is at least as good as the probability calculated with a prediction model, and may therefore not use or follow the models predictions.[18] It may thus be important to compare physicians' predictions with those of prediction models, preferably already during the development phase of a prediction model, but certainly in validation and impact studies. The EPR can be used for this purpose as well, as the physician can compare his/her own predicted probability with the predicted probability of the model that appears on the screen. Next, prediction models may not be used because they are not user-friendly[3,50]. The user-friendliness of a prediction model depends on the way a prediction model is presented; the original formu-

la is the most exact and accurate form, but may involve cumbersome calculations requiring a calculator or computer. Although sumscores, risk stratification charts or nomograms may be less precise, they certainly are more user-friendly. Moreover, simplified predictions may well be sufficiently precise for use in daily practice. However, with the EPR, the use of formulas in daily practice will become much easier, as they can be incorporated in the EPR. As a result, the complex calculations are 'hidden' in spreadsheets

Updating

Once a prediction model is implemented in daily practice via the EPR, information of new patients is constantly added to the EPR. Therefore, the EPR can be used to regularly test the performance of implemented prediction models. For example, the performance of a prediction model may decrease over time when predictor assessment has changed (e.g. due to improved imaging or measurement techniques), or when the outcome incidence has decreased due to improved care. Moreover, one might want to test whether new predictors, such as newly available diagnostic tests, improve the performance when included in the model.

By combining the information captured in the original model with the information of the newly added patients, the incorporated prediction models can be updated (Figure 1, both Research and Practice arrows). As a result, the updated models are better adjusted to the new patients and will likely perform better in future patients. Several updating methods of various complexity can be used. [21-25]

In conclusion

The EPR is used by physicians to register the diagnostic and prognostic process, primarily to aid patient care. The great advantage of the EPR is that the quality and completeness of the data closely reflects routine patient care. However, one should be realistic in using the EPR for research purposes. Information that is needed for prediction research and not so much for patient care will not automatically be collected by all physicians. In general, more impediments are experienced when a model is developed on retrospectively than prospectively collected data. There are exceptions where all potential predictors have been routinely recorded for all patients. For example, when the potential predictors are all routinely registered demographic variables (age, gender, type of insurance, etc) and the generally accepted medical data, the prediction model is just as easily developed on retrospective as prospective data.

Although patient characteristics (symptoms, signs, patient demographics, disease characteristics, test results, interventions, diagnoses and prognostic events) are ideally documented according to standardised international classifications, in practice this will not always be the case. Therefore, if one considers incorporating the use of standard classifications in an EPR, the physicians should be involved in the design phase of such EPR. Although for prediction research the registration of the documentation of symptoms and signs that arise the suspicion of a diagnosis is as important as the registration of the final diagnosis, this may be perceived as redundant and time consuming. Again, physicians involvement is required to enable success. Also for the implementation of prediction models in the EPR the intended users should be consulted. Prediction models, and the way they are implemented (directive of indicative)[5], should match the needs of their intended users.

Once a prediction model is incorporated in the EPR, information of new patients is

constantly added to the EPR facilitating regular validation and updating of the models. Regular updating enables the models to contain the latest clinical and scientific findings in prediction research. This can consistently enhance patient care.

References

1 Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993; 118(3):201-210.
2 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997; 277(6):488-494.
3 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(4):361-387.
4 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130(6):515-524.
5 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006; 144(3):201-209.
6 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.
7 Moons KG, Bots ML, Salonen JT, Elwood PC, Freire dC, Nikitin Y et al. Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? J Epidemiol Community Health 2002; 56 Suppl 1:i30-i36.
8 Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clin Chem 2004; 50(3):473-476.
9 Oudega R, Moons KG, Hoes AW. Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care. Fam Pract 2005; 22(1):86-91.
10 Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003; 56(6):501-506.
11 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10):1087-1091.
12 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12):1255-1264.
13 Harrell FE, Jr. Regression modelling strategies. Springer-Verlag, New York; 2001.
14 Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, New Jersey: John Wiley & Sons; 1987.
15 Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10):1092-1101.
16 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10(4):585-598.
17 Schafer JL. Analysis of Incomplete Multivariate Data. Chapman & Hall /CRC; 1997.
18 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(1):79-84.
19 Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005; 94(1):200-205.
20 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;

19(4):453-473.
21 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23(16):2567-2586.
22 Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med 1986; 105(4):586-591.
23 Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med 1986; 146(1):81-83.
24 Hosmer D.W., Lemeshow S. Applied logistic regression. New York: John Wiley and Sons, Inc.; 1989.
25 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999; 99(16):2098-2104.

Chapter 6

Summary

In clinical prediction research, patient characteristics, test results and disease characteristics are often combined in so-called prediction models to estimate the risk that a disease or outcome is present (diagnosis) or will occur (prognosis). In daily practice physicians often make implicit risk estimations using their clinical knowledge and experience. Prediction models are meant to make these estimations more explicit and to further aid them in their decision making. Roughly four phases can be distinguished in clinical prediction research; derivation of the prediction model, validation of the model in new subjects (and updating if necessary), assessment of its clinical impact, and the application in daily clinical practice. This thesis aims to improve methods of clinical prediction research, with a focus on the derivation, validation, updating, and application of prediction models.

In **Chapter 2.1**, we compare three methods that can handle missing predictor values when a prediction model is derived: complete case analysis, dropping the predictor with missing values and multiple imputation. We used the data from a cross sectional study that aimed to predict the presence of deep venous thrombosis (DVT), in which 804 consecutive patients presented themselves to the physician with a suspicion of DVT. The percentage of missing values in one of the predictors varied between 10% and 90%, and we applied the three methods to handle these missing values. Multiple imputation outperformed complete case analysis and dropping the variable with missing data from the analysis in terms of bias and coverage of the 90% confidence interval. The ROC area after multiple imputation was unbiased, but substantially decreased when complete case analysis was conducted or when the variable with missing data was dropped. We concluded that multiple imputation is to be preferred over dropping the variable with missing data or complete case analysis.

In **Chapter 2.2**, we used four methods to develop a model that predicts the presence or absence of Deep Venous Thrombosis (DVT); logistic regression, logistic regression with a single shrinkage factor, logistic regression with inherent shrinkage by penalised maximum likelihood estimation (PMLE), and genetic programming. In logistic regression, the selection of the predictors is based on the maximization of the log likelihood of the model. This provides the optimal fit to the data and often results in fitting noise and unstable parameter estimates. To prevent this overfitting, internal validation techniques can be applied, such as bootstrapping techniques. Instead of maximizing the log likelihood, PMLE maximizes the penalized log-likelihood, in which the maximum log likelihood of the model is adjusted by a penalty factor. Genetic programming is a novel and promising search method that may improve the selection of predictors and may lead to models with good predictive performance in new patients. The models were derived on a derivation set (1668 patients) and tested on a validation set (418 patients). The performance measures (discrimination and calibration) of the four models in the validation sets were only slightly different, and the 95% confidence intervals of the areas mostly overlapped. The choice between these derivation methods should be based on the characteristics of the data and situation at hand.

Chapter 3.1 is a review describing important aspects of validation studies and updating methods, impact analyses and the implementation of prediction models. Validation studies should consist of an adequate sample of 'different but related patients' compared to

the development study population, in which relatedness is defined as 'patients suspected of the same disease' for a diagnostic rule, and for a prognostic rule as 'patients at risk of the same event'. In validation studies, temporal, geographical, and domain validation can be distinguished. Temporal validation tests the generalisability of a prediction rule 'over time'. Geographical validation studies typically test the generalisability of a prediction rule in a patient population that is similarly defined as the development population, though in other geographical areas. Domain validation is the broadest form of validation, and tests the generalisability of a prediction model across different domains, such as patients from a different setting (primary, secondary or tertiary care), inpatients versus outpatients, patients of different age categories (e.g. adults versus adolescents or children), of a different gender, and perhaps from a different type of hospital (academic versus general hospital). In general, the potential for differences between the derivation and validation population is smallest in a temporal validation study, and largest in a domain validation study. As a result, good results on a domain validation study are considered to provide the strongest evidence that the prediction model can be generalised to new patients. Various updating methods that can be applied to improve the predictive performance are discussed. The need for impact analyses to assess the true effect of the implementation of a prediction model in daily clinical practice is discussed, and the barriers during the implementation of a model in clinical practice are considered. The review ends with remaining methodological issues in prediction research.

In **Chapter 3.2**, we modified and validated a prediction model that originally was derived to preoperatively predict the risk of severe pain in the first postoperative hour in surgical inpatients. We modified the model to enhance its use in both inpatients and outpatients, using the data of outpatients that underwent surgery in the same hospital during the same period as the inpatients. Modification of the prediction model included reclassification of the predictor 'type of surgery' and addition of interaction terms between surgical setting (ambulatory surgery: yes/no) and the other predictors. As the incidence of severe postoperative pain could be different in other patient populations, we estimated what the effect of the different incidences would be on the intercept, and presented these adjusted intercepts with the prediction model. Subsequently, we externally validated the modified model in patients that underwent surgery in another hospital later in time (temporal and geographical validation). The modified prediction model showed good calibration and reasonable discrimination in both inpatients and outpatients. By validating the modified model in patients that underwent surgery more recently and in another hospital, the model proofed to be generalisable in place and time.

In **Chapter 3.3**, we present a simple updating strategy that can improve the performance of a model in new patients. The prediction model was derived to predict the risk of severe postoperative pain in both inpatients and outpatients (Chapter 3.2). When we validated the model in patients that were treated later in time and in another hospital, the predictive performance, notably the calibration, was poor. This poor calibration was expected, as there was a difference in outcome incidence between the derivation set and the validation set. When a physician does not have information of new patients but knows that there is a difference in outcome incidence, an adjusted intercept can be used (Chapter 3.2). However, when a physician has information of new patients, we advise to validate the model

and to update it if necessary. We applied an easy updating strategy that adjusts only the intercept of the original model using the data of the new patients. This simple updating strategy already improved the predictive performance in the new patients.

In **Chapter 3.4**, we discuss several updating methods that improve the predictive performance of a model in new patients. The updating methods vary in extensiveness, which is reflected by the number of model parameters that is adjusted or re-estimated. We used a derivation set to derive the prediction model, we updated it in an updating set and subsequently tested it in new patients (test set). Five updating methods were used. Method 1 and 2 were recalibration methods, methods 3, 4 and 5 were revision methods. In method 1, only the model intercept was adjusted. In method 2, the model intercept was adjusted and the regression coefficients were multiplied with one correction factor. Method 3 tested whether the effects of some predictors were different in the updating set. In method 4 and 5, the intercept and the regression coefficients of all predictors were re-estimated in respectively the updating set and the combined derivation and updating set. Calibration of the original model was poor and substantially improved by all five updating methods, in both the updating and test set. The discrimination seemed to be improved by the revision methods, but this did not hold in the test set, in which all updated models showed a similar discrimination as the original prediction model. We concluded that simple recalibration methods improved the calibration of the original model to a similar extent as the more complex revision methods.

Chapter 4.1 focuses on the problem of missing predictor values when a prediction model is applied in daily clinical practice. We presented six strategies that handle missing values and compared the effects on the predictive performance of a prediction model that predicts the presence of deep venous thrombosis (DVT). We derived and externally validated the prediction model in respectively 1295 and 532 primary care patients. In an application set (259 patients) we mimicked three scenario's, in which an important predictor, a weak predictor and both predictors simultaneously would not be available (missing). The six strategies we used were imputation of the value zero, mean imputation, subgroup mean imputation, multiple imputation, and applying a submodel consisting of only the observed predictors, either estimated in the derivation set or by the so-called one step sweep strategy. Of the six methods, multiple imputation was best capable of improving the predictive performance of the prediction model in new patients when one or more predictor values were missing.

Chapter 5 presents an overview of the promises and pitfalls of using the electronic patient records (EPR) as a basis for prediction research to enhance patient care, and vice versa. The EPR are medical records in digital format that facilitate storages and retrieval of data on patient care. The great advantage of the EPR is that the quality and completeness of the data reflect routine patient care. Though the primary aim of the EPR is to aid patient care it creates highly attractive opportunities for research, notably with regard to diagnostic and prognostic prediction studies. We discuss that for proper derivation of a prediction model all potential predictors need to be systematically defined and recorded in the EPR. For proper validation, of course the predictors included in the final model need to be recorded. However, information that is needed for prediction research and

not so much for patient care will not automatically be collected by all physicians. Once a prediction model is incorporated in the EPR, information of new patients is constantly added to the EPR facilitating regular validation and updating of the models. The regular updating enables the models to contain the latest clinical and scientific findings in prediction research, which can consistently enhance patient care.

Chapter 6

Samenvatting

In klinisch predictieonderzoek worden patiëntkarakteristieken, testresultaten en ziektekarakteristieken vaak gecombineerd tot zogenaamde predictiemodellen om de kans op de aanwezigheid van een ziekte of uitkomst te voorspellen (diagnostiek) of het optreden ervan in de toekomst (prognostiek). Artsen maken vaak impliciete kansberekeningen op basis van hun klinische kennis en ervaring. Door middel van het toepassen van predictiemodellen wordt geprobeerd om deze kansberekeningen explicieter te maken om artsen bij te kunnen staan in hun keuzes. We kunnen grofweg vier fasen onderscheiden in predictie onderzoek, namelijk 1) de ontwikkeling van een model, 2) de validatie van het model in nieuwe patiënten (en het updaten van het model als dat nodig is), 3) het vaststellen van de impact van het model, 4) en de toepassing van het model in de dagelijkse praktijk. In dit proefschrift worden methoden beschreven die predictie onderzoek kunnen verbeteren. Hierbij ligt de nadruk op de ontwikkelingsfase, de validatie- en updatingfase, en de toepassing van de modellen.

In **hoofdstuk 2.1** vergelijken we drie methoden om missende predictorwaarden te hanteren op het moment dat een predictiemodel wordt ontwikkeld; complete case analyse, het weglaten van een predictor met missende waarden uit het model, en multipele imputatie. Voor deze vergelijking gebruiken we data uit een diagnostische studie naar het voorspellen van de aanwezigheid van diep veneuze trombose (DVT). In deze studie kwamen 804 opeenvolgende patiënten bij de huisarts met een verdenking op DVT. We creëerden missende waarden in één predictor, oplopend van 10% tot 90%, en we pasten de drie methoden toe. Multipele imputatie resulteerde in de minste bias en de beste dekking van het 90% betrouwbaarheidsinterval. De oppervlakte onder de Receiver Operating Characteristic (ROC) curve werd correct geschat na toepassing van multipele imputatie, en was te laag na toepassing van complete case analyse en als de predictor met missende waarden werd weggelaten uit het model. We concludeerden dat multipele imputatie de voorkeur zou moeten hebben boven een complete case analyse of het weglaten van de predictor met missende waarden uit het model.

In **hoofdstuk 2.2** vergelijken we vier methoden om een model te ontwikkelen dat het risico op DVT voorspelt, namelijk logistische regressie, logistische regressie waarbij wordt gecorrigeerd voor overfitting door middel van bootstrappen, logistische regressie waarbij wordt gecorrigeerd voor overfitting door middel van 'penalised maximum likelihood estimation'(PMLE), en 'genetic programming'. Bij logistische regressie is de selectie van de predictoren gebaseerd op het maximaliseren van de log likelihood van het model. Dit resulteert in een optimale fit op de data waardoor er ook regelmatig ruis wordt meegenomen en de parameters van het model mogelijk instabiel zijn. Om deze overfitting te voorkomen kunnen interne validatietechnieken worden toegepast, zoals bootstrap methoden. In plaats van het maximaliseren van de log likelihood maximaliseert PMLE de 'penalized' log likelihood van een model waarbij de maximale log likelihood gecorrigeerd wordt met behulp van een 'penalty factor'. 'Genetic programming' is een nieuwe en veelbelovende zoekstrategie die de selectie van predictoren kan verbeteren en kan leiden tot modellen met een beter voorspellend vermogen. De modellen werden ontwikkeld in een ontwikkelingsset (1668 patiënten) en getest in een validatieset (418 patiënten). Het voorspellend vermogen (calibratie en discriminatie) van de vier modellen verschilde slechts minimaal van elkaar, en de 95% betrouwbaarheidsintervallen vertoonden grote overlap. Voorkeur

voor één van de verschillende derivatiemethoden hangt af van de specifieke studiesituatie en de karakteristieken van de data.

In **hoofdstuk 3.1** geven we een overzicht van belangrijke aspecten van validatiestudies en updatingmethoden, impact analyses, en de implementatie van predictiemodellen in de praktijk. Een validatie wordt idealiter toegepast op een populatie van andere, maar soortgelijke patiënten. Met 'soortgelijke patiënten' worden patiënten bedoeld die een verdenking hebben op dezelfde ziekte (bij een diagnostisch model), of patiënten die eenzelfde ziekte kunnen ontwikkelen (bij een prognostisch model) als de patiënten uit de ontwikkelingspopulatie. Temporele, geografische en domein validaties kunnen worden onderscheiden. Een temporele validatie test de generaliseerbaarheid van een model in een andere periode. Een geografische validatie test de generaliseerbaarheid van een model in andere ziekenhuizen, instituten of in andere geografische gebieden. Een domein validatie test de generaliseerbaarheid in bijvoorbeeld patiënten uit een andere setting (eerste, tweede of derde lijn), patiënten uit andere leeftijdscategorieën (voorbeeld volwassenen versus tieners of kinderen), patiënten van de andere sekse of patiënten uit een ander type ziekenhuis (academisch versus perifeer). Meestal zijn de verschillen tussen de ontwikkelingspopulatie en de validatiepopulatie het kleinst bij een temporele validatie, en het grootst bij een domein validatie. Daarom is een goede domein validatie een beter bewijs dat een model generaliseerbaar is naar nieuwe patiënten dan een temporele validatie. Vervolgens worden verschillende updatingmethoden beschreven om het voorspellende vermogen van een model te verbeteren. Ook de noodzaak tot het uitvoeren van impact analyses om het ware effect van een model te meten wordt besproken, en de barrières die men tegenkomt tijdens het implementeren van een model. Het hoofdstuk wordt afgesloten met enkele toekomstige methodologische uitdagingen in predictieonderzoek.

In **hoofdstuk 3.2** modificeren en valideren we een predictiemodel dat de kans op ernstige pijn in het eerste uur na de operatie voorspelt. Het model was ontwikkeld voor patiënten die na de operatie minimaal één nacht in het ziekenhuis verblijven. Het model werd gemodificeerd zodat het ook toepasbaar was voor patiënten die poliklinisch werden geopereerd. Hiervoor werd de data van poliklinische patiënten gebruikt die in dezelfde periode en in hetzelfde ziekenhuis geopereerd werden. De indeling in type chirurgie werd aangepast, en interactietermen tussen chirurgische setting (poliklinisch ja/nee) en de andere predictoren werden gemodelleerd. Omdat de incidentie van ernstige pijn kan afwijken in andere patiëntenpopulaties schatten we het effect van verschillende incidenties op het intercept zodat het intercept gemakkelijk kon worden aangepast in nieuwe populaties. Vervolgens valideerden we het gemodificeerde model in patiënten die in een ander ziekenhuis en in een andere periode geopereerd werden (temporele en geografische validatie). Het gemodificeerde model had een goed calibrerend en een redelijk discriminerend vermogen. Door het model te valideren in patiënten die in een ander ziekenhuis en in een andere periode geopereerd werden bleek het model generaliseerbaar te zijn in tijd en plaats.

In **hoofdstuk 3.3** laten we een eenvoudige updating strategie zien die het voorspellend vermogen van een model kan verbeteren in nieuwe patiënten. Het model was ontwikkeld om de kans op ernstige pijn in het eerste uur na een operatie te voorspellen (hoofdstuk

3.2). Het model valideerde minder goed vooral wat betreft de calibratie, in patiënten die in een ander ziekenhuis en in een andere periode werden geopereerd. Het verschil in de incidentie van ernstige pijn tussen de ontwikkelingsset en de validatieset kan een oorzaak zijn van de slechte calibratie. Als een arts weet dat de incidentie in zijn populatie anders is, maar verder geen data van patiënten tot zijn beschikking heeft, kan het model worden toegepast met een gecorrigeerd intercept (Hoofdstuk 3.2). Wij lieten zien dat als een arts wel data van patiënten tot zijn beschikking heeft, een eenvoudige updating strategie voldoende kan zijn om het voorspellend vermogen te verbeteren.

In **hoofdstuk 3.4** laten we vijf verschillende updating methoden zien die het voorspellend vermogen van een predictiemodel kunnen verbeteren in nieuwe patiënten. De methoden variëren in complexiteit, wat weergegeven wordt door het aantal parameters van het model dat gecorrigeerd of opnieuw geschat wordt. We maakten gebruik van een ontwikkelingsset om het model te ontwikkelen, een updatingset om het model te updaten volgens de verschillende methoden, en een testset om het voorspellend vermogen van de geupdate modellen te testen. Methode 1 en 2 zijn recalibratie methoden, methode 3, 4 en 5 zijn revisie methoden. In methode 1 werd alleen het intercept aangepast. In methode 2 werd het intercept aangepast en werden de regressiecoëfficiënten vermenigvuldigd met een correctie factor. In methode 3 werd getest of het effect van bepaalde predictoren verschillend was in de updating set in vergelijking met de ontwikkelingsset. In methode 4 en 5 werden alle regressiecoëfficiënten opnieuw geschat in respectievelijk de updatingset en de gecombineerde ontwikkelingsset en updatingset. De calibratie van het originele model was onvoldoende en verbeterde door alle updating methoden, in de updatingset en de testset. De discriminatie leek te verbeteren door de revisie methoden, maar dit effect bleef uit in de testset, waarin alle methoden tot dezelfde discriminatie leidden als het originele model. We concludeerden dat de eenvoudige recalibratie methoden de calibratie in dezelfde mate verbeterden als de meer complexe revisie methoden.

In **hoofdstuk 4.1** presenteren we zes strategieën om missende predictorwaarden te hanteren op het moment dat een predictie model wordt toegepast in de praktijk. We vergeleken het effect van de zes strategieën op het voorspellend vermogen van een predictiemodel dat de aanwezigheid van DVT voorspelt. Het model werd ontwikkeld en gevalideerd in de data van respectievelijk 1295 en 532 patiënten uit de huisartspraktijk. In een toepassingsset (259 patiënten) bootsten we drie scenario's na waarin een sterke predictor, een zwakke predictor en beide predictoren tegelijk missend waren. De zes strategieën waren imputatie van de waarde nul, imputatie van het gemiddelde, imputatie van het subgroepgemiddelde, multipele imputatie, en het toepassen van een submodel met alleen de geobserveerde predictoren (ofwel opnieuw geschat in de ontwikkelingsset, ofwel geschat door middel van de 'one step sweep' strategie). Multipele imputatie bleek het best in staat om het voorspellende vermogen van het predictiemodel te verbeteren in nieuwe patiënten wanneer één of meerdere predictoren missend waren.

In **hoofdstuk 5** wordt een overzicht gegeven van de mogelijkheden en onmogelijkheden van het Elektronisch Patiënten Dossier (EPD) als basis voor predictie onderzoek om de patiëntenzorg te verbeteren, en vice versa. EPDs zijn medische statussen in een digitaal format die de opslag en het opvragen van patiëntendata faciliteren. Een groot voordeel

van het EPD is dat de kwaliteit en de volledigheid van de data de patiëntenzorg in de dagelijkse praktijk weergeeft. Hoewel het primaire doel van het EPD de verbetering van patiëntenzorg is, creëert het EPD ook mogelijkheden voor wetenschappelijk onderzoek, in het bijzonder voor diagnostisch en prognostisch predictieonderzoek. Om een model op een adequate manier te ontwikkelen dienen de potentiële predictoren systematisch gedefinieerd en geregistreerd te worden. Om een model op een adequate manier te valideren hoeven alleen de predictoren uit het model gedefinieerd en geregistreerd te worden. Informatie die niet per definitie nodig is voor de patiëntenzorg maar wel nut heeft voor wetenschappelijk onderzoek, zal echter niet automatisch geregistreerd worden door alle artsen. Zodra een predictiemodel ingebouwd is in het EPD wordt informatie van nieuwe patiënten continu toegevoegd. Hierdoor wordt het mogelijk om modellen regelmatig te valideren en te updaten. Door deze regelmatige updating kunnen de laatste klinische en wetenschappelijke bevindingen worden geïmplementeerd in de predictiemodellen, zodat de patiëntenzorg kan verbeteren.

Chapter 6

Dankwoord

Dit proefschrift is het resultaat van een samenwerking met verschillende personen. Een aantal van hen wil ik in het bijzonder bedanken voor hun bijdrage.

Prof. Dr. K.G.M. Moons en Dr. Y. Vergouwe. De discussies met jullie waren tegelijkertijd inspirerend en motiverend. Daardoor kwam ik altijd wijzer terug van een overleg. Ik heb dit promotietraject behalve als zeer leerzaam ook als erg prettig ervaren en kan terugkijken op een leuke tijd.
Beste Carl, bedankt voor het vertrouwen dat ik de afgelopen jaren heb gekregen. Hoewel ik je agenda steeds voller heb zien worden heb je altijd tijd voor me gemaakt. Het was een genoegen met je samen te werken.
Beste Yvonne, ik heb erg veel aan je gehad als dagelijks begeleider; variërend van het onder de knie krijgen van het werken met S-plus en R, tot het aanbrengen van structuur in mijn artikelen. Het was erg prettig om met iemand samen te werken die zoveel inhoudelijke kennis heeft. Ik had me geen betere co-promotor kunnen wensen.

Prof. Dr. D.E. Grobbee. Beste Rick, ik vond het prettig dat je als tweede promotor bij mijn onderzoek betrokken bleef en hebt meegewerkt aan de vorderingen van mijn proefschrift. Bedankt voor je essentiële bijdrage bij het tot stand komen van de discussie.

Prof. Dr. C.J. Kalkman. Beste Cor, dankzij jou werd mijn pijnregel aan de noodzakelijke kritische klinische blik onderworpen. Bedankt voor het meedenken in de eerste fase van mijn promotie.

Dr. A.R.T. Donders. Beste Rogier, de manier waarop je de meest ingewikkelde concepten en analyses op een simpele manier weet uit te leggen is lovenswaardig. Bedankt voor je hulp bij mijn artikelen, ik heb altijd met plezier met je samengewerkt dus ik blijf het jammer vinden dat je uit Utrecht bent vertrokken.

Prof. Dr. F.E. Jr. Harrell. Dear Frank, I appreciate your advice and comments on my manuscripts, I hope our collaboration will continue in the future. Thank you for your hospitality during my research visit to your department. It was both inspiring and fun working with you.

(Voormalige) leden van het predictieoverleg: Corné Biesheuvel, Jolanda van den Bosch, Eva Suarthana, Teus Kappen, Martijn van Eck, Peter Zuithoff, Diane Toll, Lidewij Broekhuizen, Geert-Jan Geersing, Joris de Groot en Mireille Baart. Door ieders verschillende achtergrond is het prettig discussiëren over de predictie-literatuur. Ik heb dit altijd als erg leerzaam ervaren. De frequentie van de etentjes mag van mij omhoog!

Leuke collega's zijn goud waard. Mirjam Knol, Rolf Groenwold, Annette Baas, Anne Simkens, Nadine Goessens, Marjolein Kamphuis, Corné Biesheuvel, Esther Molenaar, Diane Toll, Maud Maessen, Martijn Verheus, Nicky Peters en alle andere kamergenoten en promovendi: met collega's als jullie is het altijd gezellig op het Julius Centrum. Ik heb me dan ook erg goed vermaakt, tijdens de lunch, inloopmomenten of uitjes buiten werktijd.

Anne May en Beate Brouwer. Het was erg leuk om met jullie de laatste fase van mijn promotie in te gaan in het kleine kamertje. Het is me goed bevallen om behalve de promotieperikelen ook mijn zwangerschap, de laatste roddels en de ins en outs van de schapenwei door te nemen. Ik heb erg met jullie kunnen lachen!

Andere Jacco. Bedankt voor je hulp bij de omslag van mijn proefschrift èn voor je vriendschap.

Paranimfen Mirjam Knol en Rieneke Lanting. Mirjam, je bent altijd de eerste naar wie ik toe ga met epidemiologische vragen. Daarnaast kan ik je als persoon erg waarderen en ben ik blij dat we nog een tijd collega's blijven. Rieneke, je directheid en humor zijn een fenomenale combinatie, ik ben dan ook erg blij met onze vriendschap. Fijn dat ik ook bij dit 'afstuderen' (ik heb me er maar bij neergelegd) een beroep op je kan doen.

Anke, Annet, Douwe, Jos, Judith, Natalie en Rieneke. Het leven is mooi maar met jullie ook nog eens erg leuk! De massale uittocht naar India, Argentinië en Bennekom mag de pret dan wel tijdelijk drukken, ik ga ervan uit dat jullie spoedig tot inkeer komen en er weer gewoon in Utrecht afgesproken kan worden (Annet, jij mag deze laatste opmerkingen als niet-geschreven beschouwen omdat je voor een emigrant uitzonderlijk mobiel blijkt te zijn). Jos, jouw visie op het leven is uniek. Hoewel je het er vaker over hebt om voor langere tijd te gaan backpacken hoop ik dat je nog een hele tijd bij ons op de bank zit (en cynische opmerkingen maakt).
Hoewel een aantal van jullie al jaren grof geschut inzetten om de volgorde te beïnvloeden heb ik jullie namen wijselijk op alfabetische volgorde gezet. Natalie, om dat goed te maken begint de laatste zin met jouw naam…

Het warme nest: Ester & Rob, Steffie, José, noonk Mathieu & tant José: bedankt voor jullie gastvrijheid. Papa en mama: in warme herinnering. Lasse, Maaren, Kars en Jorre: met jullie heb ik het nog nooit over mijn proefschrift gehad, maar toch horen jullie zeker in dit dankwoord thuis. Jullie staan namelijk voor de leukste 'nevenactiviteiten' die ik me kan wensen.

Lieve Jacco, ik hoef je weinig uit te leggen over de daginvulling in het wetenschappelijk onderzoek. Bedankt voor het luisteren en voor je advies. Samen leven met jou doet mij meer dan goed.

Chapter 6

Curriculum vitae

Kristel Janssen was born in Eijsden, the Netherlands, on November 10th 1975. She graduated from the Provinciale Secundaire School in Voeren, Belgium in 1994, and started Medical Biology at the Faculty of Medicine, Utrecht University. She started her training in Nutrition and Public Health at the Agricultural University Wageningen in 1995 and graduated in 2001 on two subjects; Lifestyle and Public Health, and Epidemiology. During her training, she conducted three research projects. The first project was conducted in collaboration with the Municipal Health Service Nijmegen on lifestyle and behaviour of adolescents in Wijchen. The second project was conducted in collaboration with the Food and Agriculture Organization, Regional Office for Asia and the Pacific (FAO/RAP), Bangkok, Thailand on differences in the National Lunch Program. The third project was conducted in collaboration with the Dutch Cancer Institute (NKI) Amsterdam on menstrual cycles and subfertility.

In 2002 and 2003 she worked as a junior researcher at the department of Medical Technology Assessment of the University Medical Center Nijmegen on a research project on the quality of life and care of palliative patients, under supervision of Dr. PFM Krabbe. In February 2004 she started the work described in this thesis at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, under supervision of Prof. Dr. KGM Moons, Prof. Dr. DE Grobbee and Dr. Y Vergouwe. She obtained her Master of Science Degree in Clinical Epidemiology at the Netherlands Institute for Health Sciences, Erasmus Medical Center Rotterdam in June 2006. From September 2006 till December 2006, she conducted a research visit to Prof. Dr. DB Rubin at the Department of Statistics, Harvard University (Massachusetts), and to Prof. Dr. FE Jr Harrell at the Department of Biostatistics, Vanderbilt University Medical School (Tennessee). As from August 2007 she is working as a postdoctoral researcher at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht.