ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 826–832

Journal of Clinical Epidemiology

External validation is necessary in prediction research: A clinical example

S.E. Bleeker[a,b,c,*], H.A. Moll[a], E.W. Steyerberg[d], A.R.T. Donders[b,e], G. Derksen-Lubsen[c], D.E. Grobbee[b], K.G.M. Moons[b]

[a]*Erasmus Medical Center/ Sophia Children's Hospital Department of Pediatrics, Room Sp 1545 Dr Molewaterplein 60, 3015 GJ Rotterdam, The Netherlands*

[b]*Julius Center for General Practice and Patient Oriented Research, University Medical Center, Utrecht, The Netherlands*

[c]*Juliana Children's Hospital, Emergency Department, The Hague, The Netherlands*

[d]*Center for Clinical Decision Sciences, Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*

[e]*Center for Biostatistics, Utrecht University, Utrecht, The Netherlands*

Accepted 21 November 2002

Abstract

Background and Objective: Prediction models tend to perform better on data on which the model was constructed than on new data. This difference in performance is an indication of the optimism in the apparent performance in the derivation set. For internal model validation, bootstrapping methods are recommended to provide biascorrected estimates of model performance. Results are often accepted without sufficient regard to the importance of external validation. This report illustrates the limitations of internal validation to determine generalizability of a diagnostic prediction model to future settings.

Methods: A prediction model for the presence of serious bacterial infections in children with fever without source was derived and validated internally using bootstrap resampling techniques. Subsequently, the model was validated externally.

Results: In the derivation set ($n = 376$), nine predictors were identified. The apparent area under the receiver operating characteristic curve (95% confidence interval) of the model was 0.83 (0.78–0.87) and 0.76 (0.67–0.85) after bootstrap correction. In the validation set ($n = 179$) the performance was 0.57 (0.47–0.67).

Conclusion: For relatively small data sets, internal validation of prediction models by bootstrap techniques may not be sufficient and indicative for the model's performance in future patients. External validation is essential before implementing prediction models in clinical practice.

Keywords: Prediction models; Internal validation; Bootstrap; External validation; Logistic Regression

1. Introduction

The performance of regression models used in diagnostic and prognostic prediction research is generally better on the data set on which the model has been constructed (derivation set) compared to the performance of the same model on new data (validation set) [1–9], especially in small data sets [10,11]. To address this, several approaches have been suggested to estimate a model's optimism [3,12–15], in particular bootstrap resampling techniques. Bootstrapping, crossvalidation, and split-sampling techniques are internal validation techniques, because the performance is estimated using patients from the model's derivation set only [3,12,13]. Bootstrapping involves taking a large number of samples with replacement from the original sample. In contrast to crossvalidation or split-sample approaches, bootstrap methods are very efficient, as the entire data set is used for model development, and no new data have to be collected for validation. Moreover, it has been shown that bootstrapping provides nearly unbiased estimates of predictive accuracy that are of relatively low variance [2,16]. However, only pure sampling variability is considered with bootstrap techniques, and changes in the patient population are not [5].

External validation aims to address the accuracy of a model in patients from a different but plausibly related population, which may be defined as a selected study population representing the underlying disease domain [5,9,17]. Most reports evaluating prediction models focus on the issue of internal validity, leaving the important issue of external validity behind. We will illustrate the limitations of internal validation to determine the generalizability of a prediction

* Corresponding author. Tel.: +31-10-4636024; fax +31-10-4636685.
E-mail address: sebleeker@yahoo.com (S.E. Bleeker).

0895-4356/03/$ – see front matter
doi: 10.1016/S0895-4356(03)00207-5

model. To this aim we use a clinical example from a diagnostic study on the prediction of the presence of a serious bacterial infection in children presenting with fever without apparent source in pediatric Emergency Departments.

2. Methods

Fever without apparent source is a common diagnostic and therapeutic dilemma in pediatrics. Approximately 10 to 35% of all visits at pediatric Emergency Departments concern febrile children [18–21], and in 14 to 40% no apparent source is found after history taking and physical examination [19,22]. The underlying cause of fever varies from mild viral to serious bacterial infections, such as sepsis or meningitis [19]. Bacterial infections are reported in 3 to 15% of febrile children [21–23]. First, we developed a diagnostic prediction rule for children presenting with fever without apparent source, including variables from patient history and physical examination to distinguish the children with serious bacterial infections from those without serious bacterial infections. Then, we validated the developed prediction rule internally using bootstrap techniques (derivation study). Finally, we validated this rule externally on data from a comparable patient sample, which included more recent data and data from an extra hospital (validation study).

2.1. Derivation and internal validation study

2.1.1. Patients

This study was conducted as part of a large ongoing study on pediatric diagnostic management [24–26], and was approved by The Institutional Review Boards of both participating hospitals.

Children between 1 month and 36 months of age who were referred by the general practitioner to the Emergency Department of the Sophia Children's University Hospital Rotterdam between 1988 and 1992 for the evaluation of acute fever without apparent source were enrolled (derivation set, $n = 379$). Patient data were retrieved by means of a problem-oriented patient classification system, in which the main reason for encounter after evaluation of the general practitioner or history taking by the pediatrician is classified. For detailed description of this classification system we refer to previous reports [20,24,26]. Briefly, in this classification the category "infectious diseases" comprised (1) fever with meningeal signs, (2) fever with cough, (3) fever with micturition problems, (4) fever with vomiting and/or diarrhea, (5) fever with at least two obvious signs of an upper respiratory tract infection, (6) fever with signs or symptoms of conjunctivitis, (7) fever without apparent source. The last category was applied to children with a body temperature of 38ºC or higher and for whom classifications 1 to 6, as described above, were not applicable. Children with immune deficiencies were excluded. Because of the introduction of the *Haemophilus influenzae* type b vaccination for young infants in 1993, children with a positive isolate for *Haemophilus influenzae* were excluded.

2.1.2. Potential diagnostic determinants

Data were collected by reviewing the standardized medical records and the computer-documented hospital information system. Documented data from patient history and physical examination included information on age, gender, gestational age, body weight, body temperature, duration of fever (body temperature $\geqslant$38.0ºC), coughing, vomiting, diarrhea, micturition, intake, crying pattern, vital signs, clinical appearance, and information on ear-nose-throat, skin, and the respiratory, circulatory, and abdominal tract.

2.1.3. Reference standard

For each patient, the final diagnosis was determined either by a reference standard (cultures of blood, spinal fluid, urine, stool positive for a pathogen) or based on a consensus diagnosis [27]. Outcome diagnosis was the presence or absence of a serious bacterial infection. A serious bacterial infection was defined as the presence of bacterial meningitis, sepsis, or bacteriemia, pneumonia, urinary tract infection, bacterial gastroenteritis, osteomyelitis, or ethmoiditis [24,26]. As obtaining cultures of blood, spinal fluid, urine ,or stool was dependent on the clinical evaluation and not required by protocol for each patient, a follow-up period of 2 weeks was used as the standard for ruling out the possibility of a missed diagnosis of serious bacterial infection.

2.1.4. Data analysis

The association between potential diagnostic determinants and the presence of a serious bacterial infection was assessed using logistic regression analyses (SPSS version 9.0). Continuous variables were analyzed both on a linear and a transformed scale, for example, logarithmic or quadratic, to determine which scale was the better predictor of outcome [2]. Variables with 50% or more missing values were excluded from the analyses. Then, based on the literature and clinical practice [18,19,21–23,28–36], 57 variables were considered as candidate predictors for the analyses. Of these, variables with a univariable *P*-value of .15 or less were subsequently entered into a forward stepwise multivariable logistic regression procedure. Variables with a multivariable *P*-value of less than .10 and clinically relevant were selected as predictors of a serious bacterial infection.

Some of the 57 variables had missing values. Simple exclusion of patients with missing values on one or more of the variables commonly causes biased results and decreases statistical efficiency [37,38]. Therefore, missing values in the data were completed by single imputation using SOLAS (version 2.0). This method uses all available data to impute the missing values based on the correlation between each variable with missing values and all other variables [38].

The ability of a prediction model to discriminate between children with and without a serious bacterial infection was

quantified using the area under the receiver operating characteristic curve (ROC area) [2]. The ROC area indicates the likelihood that a patient with a serious infection has a higher predicted probability among a randomly chosen pair of patients of whom only one has a serious infection. Although encountering such a pair of patients is a rather artificial situation, the ROC area is often used as the primary criterion to quantify model performance. As a measure of overall performance we considered the explained variation (R^2) by the model using the definition proposed by Nagelkerke [39]. R^2 quantifies the explained variation on the log-likelihood scale, which is the natural scale to study performance of logistic regression models.

Internal validity of the model was determined by using bootstrap techniques [2,12]. Random bootstrap samples were drawn with replacement from the derivation set consisting of all patients (200 replications). Using S-plus (version 2000) both the univariable selection of variables with a *P*-value less than .15 and the multivariable selection of variables with a *P*-value less than .10 were repeated within each bootstrap sample. The model as estimated in the bootstrap sample was evaluated in the bootstrap sample and in the derivation set. The difference between the performance in the bootstrap sample and the performance in the derivation set was considered as an estimate of the optimism in the apparent performance in the derivation set. This difference was estimated for each bootstrap sample (200 times). The 200 differences were averaged to obtain a stable estimate of the optimism. The optimism was subtracted from the apparent performance in the derivation set to estimate the internally validated performance [2,3,16]. Furthermore, a shrinkage factor was derived from the bootstrap samples by calculating the slope of the linear predictor in the derivation set, where the linear predictor was calculated with the regression coefficients as estimated in the bootstrap sample. The shrinkage factor was used as a multiplier for the logistic regression coefficients to recalibrate the predictive model [2–4,40,41]. To estimate the 95% confidence interval (CI) around the performance measures we used the empirical distributions in the 200 bootstrap samples [12].

2.2. External validation study

2.2.1. Patients, diagnostic determinants, and reference standard

To determine generalizability of the derived prediction rule to new patients, the rule was applied to a new data set (validation set, $n = 179$). This validation set included children from a different time period and from an additional Children's Hospital from a different city. It comprised children with fever without apparent source who visited the Sophia Children's University Hospital Rotterdam between 1997 and 1998, and the Juliana Children's Hospital in The Hague in 1998. Both hospitals are large innercity teaching hospitals in The Netherlands. Other inclusion and exclusion criteria were identical to those used for the derivation set. Data collection, definitions of diagnostic determinants, and the reference standard were identical to the derivation study.

2.2.2. Data analysis

The prediction model was applied to the children in the validation set. The performance (ROC area and R^2) of the model as well as the calibration were assessed. A graphical impression of the calibration of model predictions in the validation set was obtained by plotting the observed proportions versus predicted probabilities [42]. In addition, subgroup analyses per hospital (Sophia Children's University Hospital Rotterdam and Juliana Children's Hospital) were performed. Subsequently, we hypothesized that the multivariable associations of the predictors with the outcome in the validation set would not differ from those in the derivation set. As an overall test of this hypothesis we compared the reestimated regression coefficients in the validation set with the regression coefficients from the derivation set before bootstrapping. Hereto, a logistic regression analysis was performed in the validation set including a linear predictor variable based on the coefficients from the derivation set as an offset variable. This analysis assumes the regression coefficients of the derivation set to be fixed. It is a one-sample test for the coefficients in the validation set.

3. Results

The derivation set was comprised of 376 children with fever without apparent source, and the validation set consisted of 179 children who had been referred for the same reason (three respectively zero patients were excluded because of isolation of *Haemophilus influenzae*). Except for the variable pale skin, no material differences were found in the distribution of the general characteristics and the predictors between the two sets (Table 1). A serious bacterial infection was present in 20% of the children in the derivation set and in 25% of the validation set. Of the 57 considered variables in the univariable analyses, 34 had a *P*-value of .15 or less, and 24 variables had a *P*-value of .05 or less. Table 2 shows the variables with a univariable *P*-value of .01 or less, and the results of the multivariable analysis. Strong predictors of serious bacterial infection included age above 1 year, duration of fever, changed crying pattern, nasal discharge or earache in history, ill clinical appearance, pale skin, chest-wall retractions, crepitations, and signs of pharyngitis or tonsillitis. The ROC area of this model was 0.825 (95%CI: 0.78–0.87) and the R^2 32.3% (95%CI: 15.1–49.4%). The estimated optimism by bootstrapping was 0.068 and 14.1%, reducing the ROC area and R^2 to 0.756 (95%CI: 0.66–0.86) and 18.0% (95%CI: 5.7–30.0%), respectively. The shrinkage factor for correction of the regression coefficients was 0.66 (95%CI: 0.38–0.93).

Subsequently, the model was applied to the validation set to test its predictive performance. The ROC area dropped to

Table 1
Distribution of patient characteristics in the derivation and validation set[a]

	Derivation set SCH (1988–1992) $n = 376$	Validation set SCH, JCH (1997–1998) $n = 179$
Male gender	214 (57)	91 (51)
Age (years)[b]	1.1 (0.7)	1.0 (0.8)
Weeks of gestation[b]	39.3 (2.1)	39.1 (2.3)
Duration of fever (days)[b]	2.8 (2.4)	2.8 (2.5)
Changed crying pattern (only <1 year)	139 (37)	79 (44)
Nasal discharge or earache	246 (65)	129 (72)
Ill clinical appearance	219 (58)	103 (58)
Body temperature at physical examination (ºC)[b]	39.7 (1.0)	39.8 (0.9)
Pale skin	66 (18)	15 (8)
Chest-wall retractions	18 (5)	16 (9)
Crepitations	15 (4)	7 (4)
Signs of pharyngitis or tonsillitis	159 (42)	82 (46)
Serious bacterial infection present	75 (20)	45 (25)

Abbreviations: JCH, Juliana Children's Hospital; SCH, Sophia Children's University Hospital.

[a] Values represent absolute patient numbers (percentages) unless stated otherwise.

[b] Mean (standard deviation).

0.57 (95%CI: 0.47–0.67) and the R^2 to 2.0%. Fig. 1 shows a poor calibration of the model in the validation set. The predicted probabilities of the presence of a serious bacterial infection ranged from 0.02 to 0.73 (mean 0.19). In particular, in the clinically important lower categories the predicted probabilities corresponded poorly with the observed proportions. Similar differences were found when the model was tested in the children of the Sophia Children's University Hospital Rotterdam or to the children of the Juliana Children's Hospital separately.

The poor results in the external validation were confirmed by refitting the multivariable model (i.e., reestimating the regression coefficients) on the data of the validation set. Overall, the regression coefficients in the validation set were significantly different from the derivation set, in particular, the regression coefficients of four of the nine predictors (changed crying pattern, nasal discharge or earache, chest-wall retractions and crepitations). The ROC area of this refitted model was 0.70 (95%CI: 0.61–0.79), and the R^2 16.9%.

4. Discussion

The aim of this study was to construct and validate a diagnostic prediction model to distinguish children with and without serious bacterial infections in children referred with fever without apparent source. Selected predictors in the derivation set were age above 1 year, duration of fever, changed crying pattern, nasal discharge, or earache in history, ill clinical appearance, pale skin, chest-wall retractions, crepitations, and signs of pharyngitis or tonsillitis. These results agree with other published studies [19,28,29,31,32,34–36]. The model seemed to perform well, according to common performance criteria (ROC area and R^2). Bootstrapping suggested a substantial optimism. Nonetheless, applying the model to the validation set showed a much larger decrease in predictive performance. This decrease was of such a degree that the model appeared useless for pediatric care. Hence, our study illustrates and confirms that internal validation per se is no guarantee for generalizability, and thus no substitute for external validation [5].

Various aspects need to be addressed to appreciate these results. Relative to the number of patients in the derivation set ($n = 376$) and particularly the number of events ($n = 75$),

Table 2
Derivation set: variables with univariable P-value ≤0.01 and results of multivariable analysis

Characteristic	Percentage SBI absent $n = 301$	Percentage SBI present $n = 75$	OR (95%CI)[a] Univariable	OR (95%CI)[a] Multivariable[b]
Patient history				
Age >1 year	51	57	1.3 (0.8–2.2)	0.4 (0.1–1.1)
Duration of fever (days)[c]	2.5 (2.2)	3.8 (2.8)	1.22 (1.10–1.34)	1.26 (1.12–1.41)
Changed crying pattern (only <1 year)	34	51	2.0 (1.2–3.4)	5.0 (1.7–15.0)
Nasal discharge or earache	69	51	0.5 (0.3–0.8)	0.4 (0.2–0.7)
Physical examination				
Body weight (kilograms)[c]	9.8 (2.9)	8.9 (3.4)	0.89 (0.82–0.98)	—
Ill clinical appearance	55	73	2.3 (1.3–4.0)	2.6 (1.4–5.0)
Poor peripheral circulation	10	25	3.2 (1.7–6.1)	—
Pale skin	14	33	3.2 (1.7–5.7)	2.2 (1.1–4.4)
Chest-wall retractions[d]	4	9	2.7 (1.0–7.3)	3.4 (1.1–10.1)
Crepitations	2	12	6.7 (2.3–19.5)	12.8 (3.8–42.9)
Signs of pharyngitis or tonsillitis	47	24	0.4 (0.2–0.6)	0.3 (0.2–0.6)

[a] Odds ratio (95% confidence interval).

[b] Intercept of the model was −2.29.

[c] Means (standard deviations).

[d] Univariable P-value: 0.06.

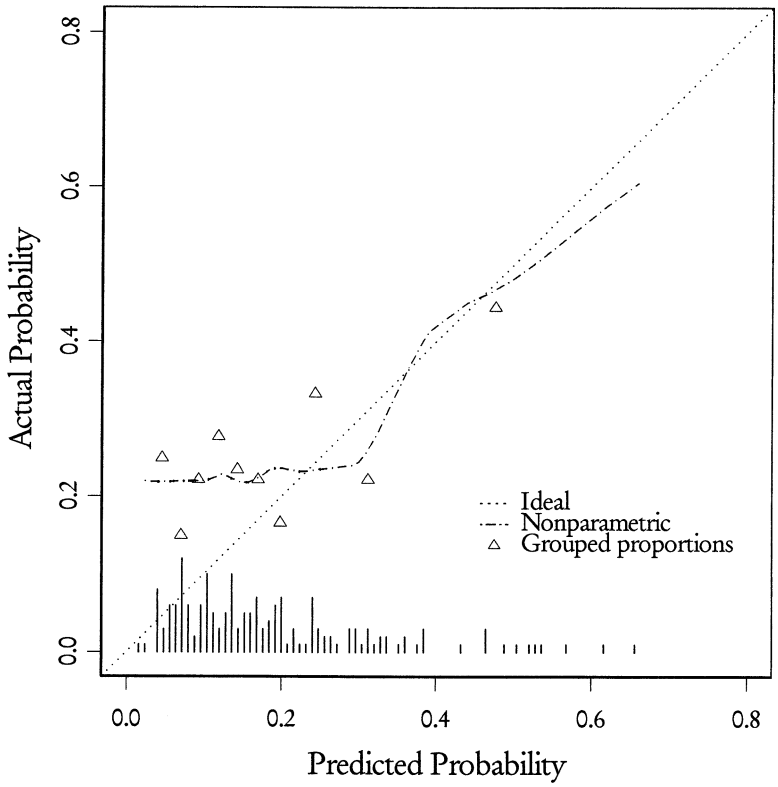

Fig. 1. Comparison of the model's predicted probabilities and observed proportions. Triangles: deciles of the predicted probabilities in the validation set (n = 179). Diagonal line: reflection of ideal situation (predicted probability = observed proportion). Dashed line: reflection of the relation non-parametrically. Lower part of the figure: histogram of the predicted probabilities.

we considered a large number of candidate predictors. Moreover, we used univariable and multivariable selection methods to develop our final prediction model. These methods are known to imply multiple testing, underestimation of standard errors and P-values, a limited power to select diagnostically important predictors, and an unstable selection of predictors, resulting in overfitting of the final prediction model [1,10,11,16]. Even though the entire variable selection process was repeated in every bootstrap sample, it might be that the estimated performance after bootstrapping was still too optimistic and the final model somewhat overfitted. A possibility to overcome this problem is to reduce the number of candidate predictors (e.g., using variable clustering methods and factor analysis) [1,2]. Furthermore, we only applied single imputation instead of the statistically more appropriate multiple imputation technique [37]. Moreover, the single imputation was only performed in the derivation set and was not repeated in the bootstrap procedure.

Apart from these statistical drawbacks, we have tried to discover other reasons of this poor validation. Information bias seems improbable, because all data of both the derivation and the validation set were collected by the same persons and before any analysis had been started.

Selection bias also seems unlikely, because the same in- and exclusion criteria were used, and Table 1 showed no major differences between the two data sets. It is generally recommended to compare the overall characteristics of the derivation and validation set before applying an internally validated prediction model to a new patient population [6,8]. These characteristics include the definition and distribution of the predictors included in the prediction model and general aspects such as the selection of patients (e.g., referral pattern). If no important differences are found, the validation set is usually considered to provide a comparable population. A similar performance as in the derivation set should then be found in the validation set.

In our study, all these aspects were comparable but did not at all assure good performance. There was indeed a slight, statistically nonsignificant, difference in frequency of serious bacterial infection between the derivation set (20%) and the validation set (25%), reflecting a difference in baseline risk (i.e., the intercept of the model). But this cannot have affected the discriminative performance substantially, because the prevalence does not directly influence estimation of the ROC area. The poor validation of our prediction model could, nevertheless, partly be explained by the fact that our inclusion criteria were rather broad, particularly in terms of age and referral patterns. Even though the two data sets used the exact same inclusion criteria, it might well be that some important predictors that were not included in the prediction model but may interact with the included predictors were differently present in the validation set.

We therefore performed some additional analyses. First, we performed a subgroup analysis per hospital. No materially different results were found. In addition, we reestimated the prediction model on the validation set to obtain insight in the maximally achievable performance of the model in the validation set. The ROC area of 0.70 and R^2 of 17% suggest that certain predictors in the original model still have some value in the validation set, although the magnitude of the associations (regression coefficients) has changed. Indeed, when exploring the validation set in further detail, the distribution of the signs and symptoms across children with and without a serious bacterial infection in the validation set appeared to be different from the derivation set. This suggests a difference in patient population, which was not exhibited by the comparison of the overall distributions (see Table 1). Such change in patient population is less likely when researchers use a more strict definition of the patient population from which a prediction model is derived. A disadvantage of the latter, however, is that the applicability of such model will decrease as well. In our study, the change in patient population might be the effect of a change in referral pattern by general practitioners, probably influenced by the introduction of the *Haemophilus influenzae* type b vaccination (April 1993) for young infants. Besides, it would be reasonable to consider the effect of vaccination on the generalizability as minor, because the occurrence of *Haemophilus influenzae* in the isolates was not appreciably changed (3 before and 0 after the introduction of the vaccine), and as those cases were excluded from the derivation set.

A final, although unlikely, explanation is that the poor validation is the result of an unfortunate data set ("bad luck"). However, we cannot exclude this contention.

Despite all above-mentioned reasons, we believe that our main conclusion that internal validation per se is no guarantee for generalizability, and thus no substitute for external validation, still stands. Although external validation is commonly needed, results of external validations are also not always unambiguous and trustworthy. External validation may require a substantial sample size to provide sufficient power to find similar performance [5]. The lack of generalizability of our model to future groups of children supports the view that clinical guidelines may not be durable with time and must be updated regularly [43].

In conclusion, after internal validation by bootstrapping of a diagnostic prediction model for serious bacterial infection in children with fever without apparent source, we had good expectations with respect to the performance of the model in the validation set. However, results from external validation showed an unexpected poor performance. This suggests that internal validation of prediction models may not be sufficient for relatively small data sets, and that external validation is necessary before implementing prediction models in clinical practice.

Acknowledgment

We gratefully acknowledge Wilfried de Jong, and Femke Mineur, medical students, for support in data collection. The Health Care Insurance Counsel of The Netherlands financially supported this project.

References

[1] Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 1985;69:1071–7.
[2] 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.
[3] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983;78:316–31.
[4] van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25.
[5] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
[6] Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313: 793–9.
[7] Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology 1999;10:276–81.
[8] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488–94.
[9] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA 2000;284:79–84.
[10] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.
[11] 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: 1059–79.
[12] Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall; 1993.
[13] Efron B, Tibshirani R. Improvements on cross-validation: the.632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.
[14] Picard RR, Berk KN. Data splitting. Am Stat 1990;44:140–7.
[15] Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. Journal of the Royal Statistical Society Series C: Applied Statistics 1999;48:313–29.
[16] Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models. Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.
[17] Knottnerus JA. Prediction rules: statistical reproducibility and clinical similarity. Med Decis Making 1992;12:286–7.
[18] McCarthy PL. Fever. Pediatr Rev 1998;19:401–8.
[19] Baraff LJ, Lee SI. Fever without source: management of children 3 to 36 months of age. Pediatr Infect Dis J 1992;11:146–51.
[20] Steensel-Moll HA van, Jongkind CJ, Aarsen RSR, Goede-Bolder A de, Dekker A, Suijlekom-Smit LWA van, et al. A problem-oriented patient classification system for general pediatrics II. Tijdschr Kindergeneesk 1996;64:99–104.
[21] Kuppermann N, Fleisher GR, Jaffe DM. Predictors of occult pneumococcal bacteremia in young febrile children. Ann Emerg Med 1998;31:679–87.
[22] Baker MD, Bell LM, Avner JR. Outpatient management without antibiotics of fever in selected infants [discussion: N Engl J Med 1994;330:939–40]. N Engl J Med 1993;329:1437–41.

[23] Baraff LJ, Oslund SA, Schriger DL, Stephen ML. Probability of bacterial infections in febrile infants less than three months of age: a meta-analysis. Pediatr Infect Dis J 1992;11:257–64.

[24] Bleeker SE, Moons KGM, Derksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. Acta Paediatr 2001;90:1226–32.

[25] Oostenbrink R, Moons KGM, Donders ART, Grobbee DE, Moll HA. Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. Acta Paediatr 2001;90:611–7.

[26] Oostenbrink R, Moons KGM, Theunissen CC, Derksen-Lubsen G, Grobbee DE, Moll HA. Signs of meningeal irritation at the emergency department: how often bacterial meningitis? Pediatr Emerg Care 2001;17:161–4.

[27] Weller SC, Mann NC. Assessing rater performance without a "gold standard" using consensus theory. Med Decis Making 1997;17:71–9.

[28] McCarthy PL, Lembo RM, Baron MA, Fink HD, Cicchetti DV. Predictive value of abnormal physical examination findings in ill-appearing and well-appearing febrile children. Pediatrics 1985;76: 167–71.

[29] McCarthy PL, Lembo RM, Fink HD, Baron MA, Cicchetti DV. Observation, history, and physical examination in diagnosis of serious illnesses in febrile children less than or equal to 24 months. J Pediatr 1987;110:26–30.

[30] Kramer MS, Lane DA, Mills EL. Should blood cultures be obtained in the evaluation of young febrile children without evident focus of bacterial infection? A decision analysis of diagnostic management strategies. Pediatrics 1989;84:18–27.

[31] Baraff LJ, Bass JW, Fleisher GR, Klein JO, McCracken GH, Powell KR, Schriger DL. Practice guideline for the management of infants and children 0 to 36 months of age with fever without source. Pediatrics 1993;92:1–12.

[32] Dagan R, Powell K, Hall C, Menegus M. Identification of infants unlikely to have serious bacterial infection although hospitalized for suspected sepsis. J Pediatr 1985;107:855–60.

[33] Isaacman DJ, Shults J, Gross TK, Davis PH, Harper M. Predictors of bacteremia in febrile children 3 to 36 months of age. Pediatrics 2000;106:977–82.

[34] Hewson PH, Humphries SM, Roberton DM, McNamara JM, Robinson MJ. Markers of serious illness in infants under 6 months old presenting to a children's hospital. Arch Dis Child 1990;65:750–6.

[35] Berger RM, Berger MY, van Steensel-Moll HA, Dzoljic-Danilovic G, Derksen-Lubsen G. A predictive model to estimate the risk of serious bacterial infections in febrile infants. Eur J Pediatr 1996;155: 468–73.

[36] Teach SJ, Fleisher GR. Duration of fever and its relationship to bacteremia in febrile outpatients three to 36 months old. The Occult Bacteremia Study Group. Pediatr Emerg Care 1997;13:317–9.

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

[38] Little RA. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.

[39] Nagelkerke N. A note on the general definition of the coefficient of determination. Biometrika 1991;78:691–2.

[40] van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Stat Neerl 2001;55:17–34.

[41] Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. Stat Neerl 2001;55:76–88.

[42] Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis, II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med 1978;17: 227–37.

[43] Shekelle PG, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? BMJ 2001;323:155–7.