

A simple method to adjust clinical prediction models to local circumstances

Une méthode simple pour ajuster les modèles de prédiction clinique aux circonstances locales

Kristel J. M. Janssen, PhD · Yvonne Vergouwe, PhD ·
Cor J. Kalkman, PhD, MD · Diederick E. Grobbee, PhD, MD ·
Karel G. M. Moons, PhD

Received: 3 July 2008 / Revised: 12 November 2008 / Accepted: 24 November 2008 / Published online: 7 February 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract

Introduction Clinical prediction models estimate the risk of having or developing a particular outcome or disease. Researchers often develop a new model when a previously developed model is validated and the performance is poor. However, the model can be adjusted (updated) using the new data. The updated model is then based on both the development and validation data. We show how a simple updating method may suffice to update a clinical prediction model.

Methods A prediction model that preoperatively predicts the risk of severe postoperative pain was developed with multivariable logistic regression from the data of 1944 surgical patients in the Academic Medical Center Amsterdam, the Netherlands. We studied the predictive performance of the model in 1,035 new patients, scheduled for surgery at a later time in the University Medical Center Utrecht, the Netherlands. We assessed the calibration (agreement between predicted risks and the observed frequencies of an outcome) and discrimination (ability of the

model to distinguish between patients with and without postoperative pain). When the incidence of the outcome is different, all predicted risks may be systematically over- or underestimated. Hence, the intercept of the model can be adjusted (updating).

Results The predicted risks were systematically higher than the observed frequencies, corresponding to a difference in the incidence of postoperative pain between the development (62%) and validation set (36%). The updated model resulted in better calibration.

Discussion When a clinical prediction model in new patients does not show adequate performance, an alternative to developing a new model is to update the prediction model with new data. The updated model will be based on more patient data, and may yield better risk estimates.

Résumé

Introduction Les modèles de prédiction clinique évaluent le risque de présenter ou de manifester un devenir ou une maladie en particulier. Les chercheurs élaborent souvent un nouveau modèle lorsqu'un modèle élaboré précédemment est validé mais que ses performances sont peu concluantes. Toutefois, un modèle peut être ajusté (mis à jour) aux nouvelles données. Le modèle mis à jour est ensuite basé aussi bien sur les données d'élaboration que de validation. Nous montrons comment une méthode simple de mise à jour peut suffire à mettre à jour un modèle de prédiction clinique.

Méthode Un modèle de prédiction qui prédit avant l'opération le risque de douleur postopératoire grave a été élaboré à l'aide de la méthode de régression logistique multivariée appliquée aux données de 1944 patients chirurgicaux du Centre médical universitaire d'Amsterdam, aux Pays-Bas. Nous avons étudié la performance

Y. Vergouwe, C. J. Kalkman, D. E. Grobbee and K. G. M. Moons are contributed equally to this work.

K. J. M. Janssen, PhD (✉) · Y. Vergouwe, PhD ·
D. E. Grobbee, PhD, MD · K. G. M. Moons, PhD
Julius Center for Health Sciences and Primary Care, University
Medical Center Utrecht, Str 6.131, P.O. Box 85500, 3508 GA
Utrecht, The Netherlands
e-mail: K.J.M.Janssen@umcutrecht.nl

C. J. Kalkman, PhD, MD
Department of Anesthesiology, Division of Perioperative Care
and Emergency Medicine, University Medical Center Utrecht,
Utrecht, The Netherlands

prédictive du modèle chez 1 035 nouveaux patients qui devaient subir une chirurgie plus tard au Centre médical universitaire d'Utrecht, aux Pays-Bas. Nous avons évalué le calibrage (accord entre les risques prédits et les fréquences observées d'un devenir) et la discrimination (capacité du modèle de distinguer entre les patients avec ou sans douleurs postopératoires). Lorsque l'incidence du devenir est différente, tous les risques prédits peuvent être systématiquement sur- ou sous-estimés. Ainsi, le point d'intersection du modèle peut être ajusté (mise à jour).

Résultats *Les risques prédits étaient systématiquement plus élevés que les fréquences observées, ce qui correspond à une différence de l'incidence des douleurs postopératoires entre les données d'élaboration (62 %) et celles de validation (36 %). Le modèle mis à jour a généré un meilleur calibrage.*

Conclusion *Lorsqu'un modèle de prédiction clinique chez de nouveaux patients ne génère pas une performance adaptée, une alternative à l'élaboration d'un nouveau modèle consiste en la mise à jour du modèle de prédiction avec de nouvelles données. Le modèle mis à jour sera basé sur davantage de données patients, et pourrait donner de meilleures estimations des risques.*

Clinical prediction models or risk scores are developed to estimate or predict a patient's risk of having (diagnosis) or developing (prognosis) a particular outcome or disease. Such models have become increasingly popular in anesthesiology, critical care, and surgery.¹⁻³ The best known risk score is undoubtedly the Apgar score.⁴ Others include the Acute Physiology and Chronic Health Evaluation (APACHE) score,⁵ Simplified Acute Physiology Score (SAPS),⁶ Framingham risk score,⁷ Ottawa ankle rule,⁸ and risk scores for predicting postoperative vomiting⁹ and pain.¹⁰ Any prediction model tends to show optimistic predictive accuracy in the data from which it was developed. A simple Medline search using a suggested search strategy¹¹ revealed numerous examples of prediction models showing lower accuracy when applied in new subjects across all medical domains.¹²⁻¹⁹ The decrease in accuracy varied, but was often large enough to adversely affect patient management and outcome.

Hence, it is widely recommended that any prediction model should first be validated in new subjects before application in practice.^{1-3,18,20,21} New subjects can be used at a later period from the same institute (temporal validation); from another institute, city, or country (geographical validation); or from another level of care, e.g., primary vs. secondary care (transmural validation).

While the number of studies developing prediction models is sharply increasing, a smaller number of prediction models have been validated.³ Researchers frequently use their data set simply to develop their 'own' prediction model, without first validating existing models. If researchers do validate existing models and discover poor performance for their data or setting, they often proceed to develop a new model by re-estimating the predictor-outcome associations or even by repeating the entire selection of important predictors. For example, there are over 60 published models aiming to predict outcome after breast cancer²² and about 25 for predicting long-term outcome in neurotrauma patients.²³ This practice is problematic for several reasons. First, developing a different model per time period, hospital, country, level of care, etc., makes prediction research particularistic and non-scientific. Second, prior knowledge is not used optimally, i.e., predictive information captured in the original model is neglected. Finally, validation studies commonly include fewer patients than the corresponding development study, making the new model more subject to limitation and, thus, even less comprehensive than the original model.

The principle of using prior knowledge has been recognized in etiologic and intervention research where meta-analyses are common. But prior knowledge can also be effectively used in prediction research. When a prediction model performs inadequately in another population or setting, it has been shown that the model can often be 'updated' (adjusted) using the new data to improve its performance in that population.^{24,25} Such an updated model is based on both the development and the validation data. Unfortunately, these updating methods are seldom used in applied clinical research. Several updating methods can be distinguished by the extent to which they vary in their comprehensiveness, as is reflected by the number of variables that are adjusted or re-estimated. In the simplest updating method, only one variable of the original prediction model is adjusted, while in the more extensive methods, the effects of several predictors are adjusted or additional variables are considered.

Given the considerable increase in the number of published prediction models across all medical domains (the number will increase with the introduction of electronic patient records),²⁶⁻²⁸ we thought it important to re-emphasize how a simple updating method can effectively adjust a prediction model to local circumstances. The prediction model described herein was developed to predict severe postoperative pain; however, the methodology can be applied to many types of prediction models across medical domains. The following is an example of a

situation where a difference in the primary outcome leads to reduced performance.

Methods

Patients

Moderate to severe acute postoperative pain occurs frequently after surgery. Incidences of up to 50% in inpatients and 40% in outpatients (patient who undergo ambulatory surgery) have been reported.^{29–31} Risk-based prophylactic treatment could reduce the frequency of postoperative pain.

A prediction model that preoperatively predicts the risk of severe postoperative pain was developed with multivariable logistic regression. The model was thoroughly developed from the data of 1,944 surgical patients selected in the Academic Medical Center, Amsterdam, the Netherlands (development set). That data set has been reported previously.^{10–32} Severe acute postoperative pain (herein named ‘postoperative pain’) was defined as a score ≥ 6 at a numerical rating scale (0 indicates no pain at all, and 10 indicates the most severe pain imaginable), which occurred at least once within the first hour after surgery. The prediction model is presented as an original regression formula (Box 1) and as an easy-to-use score chart (Fig. 1).

Fig. 1 Score chart to predict the risk of severe acute postoperative pain for inpatients and outpatients. The scores per predictor were derived by multiplying the regression coefficients by 5 and rounding to the nearest integer. A sum score can be calculated for each patient by adding the scores that correlate to the patient’s characteristics. The total sum score can be linked to the patient’s individual risk using the box in the lower part. Consider, for example, an inpatient setting (intercept = -0.42 , corresponding score = 0), a female patient ($\beta = -0.004$, score = 0) of age 64 ($\beta = -0.009 * 64 = -0.576$, score = -3), with a preoperative pain score of 7 ($\beta = 0.11 * 7 = 0.77$, score = 4), who is scheduled for a high pain procedure ($\beta = 1.05$, score = 5) with a small expected incision size ($\beta = 0$, score = 0), who has a preoperative anxiety score of 16 ($\beta = 0.05 * 16 = 0.8$, score = 4), and a preoperative need for information score of 4 ($\beta = -0.05 * 4 = -0.20$, score = -1). The intercept plus the regression coefficients times the predictor values total 1.42 using the formula in Box 1, yielding a predicted risk of pain of $1/(1 + e^{-1.42}) = 80\%$. The total score is 9, which results in a risk of postoperative pain of 81%. Reprinted with permission from Janssen et al.³²

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 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

Box 1 Original regression model to predict the risk of severe acute postoperative pain for inpatients and outpatients, where -0.42 is the intercept, and the other numbers are the regression coefficients (β) of each predictor or interaction term. Female gender, types of surgery, expected incision size ≥ 10 cm, and ambulatory surgery equal 1, if true, and 0, otherwise. Separate regression coefficients (β) for gender and type of surgery were estimated for inpatients and outpatients, as the effect of these predictors differed across in- and outpatients (interaction terms)

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

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

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

We studied the predictive performance of the model in 1,035 new patients (validation set), to test whether or not the model could be generalized across time and place. This data set has been presented and analyzed elsewhere.³² As a minimum of 100 events are required to detect changes in the predictive performance between two sets, our validation set was large enough for this purpose.³³ The patients in our 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), i.e., temporal and geographical validation. The study was approved by the institutional medical ethics committee, and all patients gave written informed consent for their participation.

Statistical analyses

We considered two aspects of the performance of the prediction model in the validation set, i.e., calibration and discrimination.

Calibration

Calibration is the agreement between the risks predicted by the model and the observed frequencies of an outcome. It can be graphically assessed with a calibration plot, with the predicted probabilities representing the independent variable and the observed frequencies representing the dependent variable. The plot ideally rests exactly on the 45° line, implying that the predicted risks are equal to the observed frequencies. However, when the incidence of the outcome is lower in the validation set, all predicted risks may be

systematically overestimated. In that situation, the intercept (that reflects the risk of the outcome not explained by the covariates) of a prediction model can easily be adjusted, such that the mean predicted risk equals the observed incidence in the validation set.³⁴ This modification is called ‘updating’, which means that the model is adjusted to the new circumstances and combines the information captured in the original model with the information (lower outcome incidence) of the new patients.^{24,25}

The correction factor for the intercept is estimated in the validation set and is based on the mean predicted risk and the incidence in the validation set.

$$\text{Correction factor} = \ln \left(\frac{\text{incidence in validation set}}{1 - \text{incidence in validation set}} \right) / \left(\frac{\text{mean predicted risk in validation set}}{1 - \text{mean predicted risk in validation set}} \right)$$

This correction factor equals the natural logarithm of the odds ratio of the mean observed incidence and the mean predicted risk. The correction factor simply needs to be added to the intercept of the original model (Box 1) when the model is applied to the new patients. Consequently, in the simplified score chart, the new intercept also needs to be adjusted to an easy to use number.

Discrimination

Discrimination is the ability of the model to distinguish the patients with postoperative pain from patients without postoperative pain and is quantified with the area under the

Table 1 Distribution of the characteristics of patients who underwent surgery in the Amsterdam cohort between April 1997 and January 1999 and patients who underwent surgery in the Utrecht cohort between February and December 2004; % (*n*) unless stated otherwise

Patient characteristics	Amsterdam cohort 1997–1999 (<i>n</i> = 1944)	Utrecht cohort 2004 (<i>n</i> = 1035)
Female gender	57 (1110)	58 (596)
Age (years) ^a	43 (15)	47 (16)
Preoperative pain score (NRS) ^a	3.0 (2.7)	3.4 (2.9)
Type of surgery		
Lowest expected incidence of pain	4 (72)	9 (92)
Low expected incidence of pain	30 (590)	29 (297)
Medium expected incidence of pain	18 (348)	19 (200)
High expected incidence of pain	35 (671)	33 (339)
Highest expected incidence of pain	14 (263)	10 (107)
Ambulatory surgery	28 (549)	43 (444)
Expected incision size ≥ 10 cm	37 (724)	7 (67)
APAIS anxiety score ^a	9.4 (4.0)	9.5 (3.9)
APAIS need for information ^a	6.5 (2.2)	6.2 (2.3)
Severe acute postoperative pain, NRS ≥ 6	62 (1205)	36 (374)

NRS numeric rating scale; APAIS Amsterdam preoperative anxiety and information scale. Reprinted with permission from Janssen et al.³²

^a Mean (standard deviation)

receiver-operator characteristic (ROC) curve (AUC). An AUC ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination). The AUC of a prediction model is measured after ranking all subjects in a data set based on their predicted risk and estimating the extent to which these predicted risks are different between subjects, with and without the outcome. The AUC is a rank order statistic.^{35,36}

Results

One-third (36%) of the patients in the validation set reported severe pain, compared to 62% of the patients in the development set (Table 1). The distribution of most predictors was similar in the two data sets, although, compared to patients in the development set, patients in the validation set were slightly older (47 vs. 43 years, respectively), had ambulatory surgery more often (43% vs. 28%, respectively) and had a lower incidence of large surgical incisions (7% vs. 37%, respectively).

Calibration

Prediction models usually show good calibration in the development set, which was also the case for the postoperative pain model (Fig. 2a). However, the model showed insufficient calibration when tested in the patients of the validation set (Fig. 2b). Predicted risks were systematically higher than observed frequencies. The question arises as to how this happened.

The systematically overestimated risk in the validation set of the prediction model corresponds to the difference in

incidence of postoperative pain between the development set (62%) and the validation set (36%). Since the incidence was 36%, and the mean predicted risk of the patients in the validation set was 57%, the correction factor was:

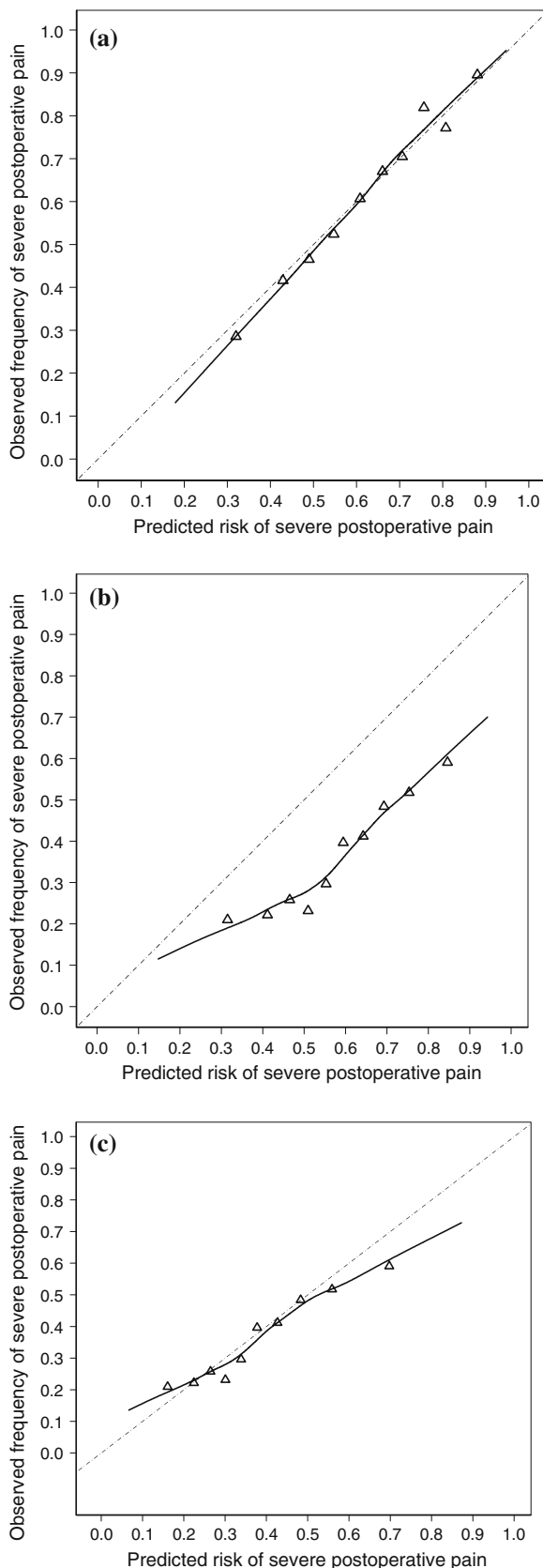
$$\ln\left(\frac{0.361}{1 - 0.361} / \frac{0.577}{1 - 0.577}\right) = \ln(0.414) = -0.89$$

The correction factor was added to the intercept of the original model (Box 1), which yields the (new) intercept of $-0.42 - 0.89 = -1.31$, that should be used when applying the model to the patients in the validation set. Consequently, in the simplified score chart, the new intercept also needed to be adjusted to an easy-to-use number. As the scores in the score chart were equal to the regression coefficients multiplied by 5 and rounded to the nearest integer (Fig. 1), the new intercept was adjusted to -7 ($-1.31 * 5 = -6.55 = -7$). For example, for the female patient discussed in the legend of Fig. 1, the adjusted prediction model summed up to 0.53, yielding a risk of $1/(1 + e^{-0.53}) = 63%$ (vs. 80% before updating) (Fig. 1). Using the simplified score, the total sum score was 2, which leads to a risk of postoperative pain of 60%.

The calibration plot of the updated model in the validation set is shown in Fig. 2c. As expected, the updated prediction model resulted in lower predicted risks and a calibration line that was much closer to the ideal line.

Discrimination

The AUC of the original prediction model (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. Adjustment of the intercept, i.e., adding or subtracting a fixed value for each



◀ **Fig. 2** Calibration line of the original prediction model in the development set (a), in the validation set (b), and the calibration line of the original prediction model 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, where the predicted risks equal the observed frequencies of severe postoperative pain. Reprinted with permission from Janssen et al.³²

subject, did not change the ranking of the predicted risks of the subjects. Thus, the AUC of the model was unaltered, and the AUC of the updated model was also 0.65 in the validation set.

Discussion

It is important to improve the scientific approaches for evaluating the generality of prediction models, since, on the one hand, we observe an increased attention for prediction models in the literature and in clinical practice, and, on the other hand, we often encounter poor accuracy of a model in new subjects. Therefore, developing and improving methods for validation and updating of prediction models will be relevant to many clinical domains dealing with diagnosis and prognosis of patients. If a model in new patients does not show sufficient performance at the outset, an alternative to developing a new model would initially be to adjust or update the previously developed prediction model(s) with the new data so as to improve its calibration and/or discrimination, provided that the initial model was appropriately developed.^{24,25,37–39} The updated model should be based on additional patient data yielding better risk estimates and should be easily transferred to other, yet untested, populations.

In our study, the model showed disappointing calibration due to a difference in outcome incidence. It must be noted that calibration would not be adversely affected if the difference in incidence were due only to differences in patient characteristics (predictors) included in the model. For example, if the lower incidence in the validation set was the result of a larger proportion of older patients (who experience postoperative pain less frequently, as reflected by a regression coefficient of -0.009 per year) or a larger proportion of patients experiencing a type of low pain surgery (regression coefficient of 0.50 compared to 1.72 for highest pain surgery), the (mean) predicted risks in the validation set would also be lower, thus closer to the observed frequencies. Accordingly, the calibration plot of the model in the validation set would be comparable to Fig. 2a. However, the mean predicted risk in the validation

set of the model (57.7%) was substantially higher than the observed frequency (36.1%). Hence, the lower postoperative pain incidence in the validation set and the overestimation by the model cannot be explained by the predictors; therefore, they must be the result of characteristics that were not included in the model. For example, the hospital used for the development study may have used less aggressive pain treatment than the validation hospital. Note that the presented correction factor is not applicable when the outcome frequency is extremely low (below 5%) or extremely high (above 95%). For extreme outcome frequencies, the correction factor can be estimated by fitting a logistic regression model with the linear predictor of the prediction model (as offset) as the only covariate in the new patients. For further elaboration, we refer to the literature.^{24,25}

Adjustment may vary from simply updating the intercept (or constant) of the model for differences in outcome frequency, as we illustrated above; to adjusting the relative weights of the predictors in the model, in case the associations of the predictors are different in the new population; to adding new predictors, in case an important predictor was overlooked. We showed that a simple intercept adjustment can greatly improve the performance of a prediction model in new patients. However, such adjustment only improves the calibration of the model. Usually more rigorous adjustments are needed to enhance a model's discrimination, for instance, adding (previously missed) important predictors to the model.^{24,25} It is recommended, however, that a considerably updated model be again validated in other populations. Furthermore, we stress that updating methods require the original regression formula of the prediction model, as exhibited in the Box 1. This means that researchers developing prediction models should not only present the simplified risk score of their model, as frequently occurs, but should also present the underlying regression model. We refer to the literature for a more comprehensive discussion of available methods for updating prediction models.^{24,25}

It would be useful if we could describe the calibration plots that call for different updating methods. Unfortunately, it is not as straightforward as it may seem, since, for example, it is not possible, based on the calibration plot, to choose between the updating method where additional predictors are considered, and the updating method where a predictor that is already in the model is re-estimated. When there is only a difference in incidence, as in our clinical example, the calibration plot may indicate the method of updating. However, this updating method cannot improve the discrimination. Therefore, the researcher must consider whether the discrimination and/or the calibration needs to be improved, whether a change in measurement may have caused a decreased discrimination and/or calibration, and

whether there are potential predictors available that may improve the discrimination and/or calibration. Therefore, it is not possible to recommend the type of updating purely based on the calibration plot.

Recent methodological advances in prediction research may see future prediction models being continuously validated and updated, while quantitatively maintaining all available evidence. The extent to which this process of model validation and updating must be pursued before clinical application is justified is yet unknown and a topic for further research. Although our paper relates to the preoperative prediction of severe postoperative pain, it may serve as a practicable contribution to improving the validation and use of prediction models in other medical domains, including perioperative emergency medicine and surgery.

Conflicts of interest None declared.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. *Randolph AG, Guyatt GH, Calvin JE, Doig DVM, Richardson WS.* Understanding articles describing clinical prediction tools. *Evidence Based Medicine in Critical Care Group. Crit Care Med* 1998; 26: 1603–12.
2. *Laupacis A, Sekar N, Stiell IG.* Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997; 277: 488–94.
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. *Apgar V.* A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953; 32: 260–7.
5. *Knaus WA, Wagner DP, Draper EA, et al.* The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100: 1619–36.
6. *Le Gall JR, Lemeshow S, Saulnier F.* A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; 270: 2957–63.
7. *Kannel WB, McGee D, Gordon T.* A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 1976; 38: 46–51.
8. *Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR.* A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med* 1992; 21: 384–90.
9. *Apfel CC, Laara E, Koivuranta M, Greim CA, Roewer N.* A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999; 91: 693–700.
10. *Kalkman CJ, Visser K, Moen J, Bonsel GJ, Grobbee DE, Moons KG.* Preoperative prediction of severe postoperative pain. *Pain* 2003; 105: 415–23.
11. *Ingui BJ, Rogers MA.* Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001; 8: 391–7.

12. Tobin K, Stomel R, Harber D, Karavite D, Sievers J, Eagle K. Validation in a community hospital setting of a clinical rule to predict preserved left ventricular ejection fraction in patients after myocardial infarction. *Arch Intern Med* 1999; 159: 353–7.
13. Sanson B, Lijmer JG, Mac Gillavry MR, Turkstra F, Prins MH, Buller HR. Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism. ANTELOPE-Study Group. *Thromb Haemost* 2000; 83: 199–203.
14. Fortescue EB, Kahn K, Bates DW. Prediction rules for complications in coronary bypass surgery: a comparison and methodological critique. *Med Care* 2000; 38: 820–35.
15. Orford JL, Sesso HD, Stedman M, Gagnon D, Vokonas P, Gaziano JM. A comparison of the Framingham and European Society of Cardiology coronary heart disease risk prediction models in the normative aging study. *Am Heart J* 2002; 144: 95–100.
16. Suistomaa M, Niskanen M, Kari A, Hynynen M, Takala J. Customized prediction models based on APACHE II and SAPS II scores in patients with prolonged length of stay in the ICU. *Intensive Care Med* 2002; 28: 479–85.
17. Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med* 2003; 29: 249–56.
18. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003; 56: 826–32.
19. Oudega R, Hoeks AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med* 2005; 143: 100–7.
20. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; 130: 515–24.
21. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; 19: 453–73.
22. 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: Informa Healthcare; 2007. p. 11–25.
23. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006; 6: 38.
24. 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.
25. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008; 61: 76–86.
26. James BC. Making it easy to do it right. *N Engl J Med* 2001; 345: 991–3.
27. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005; 330: 765.
28. Hippisley Cox J, Pringle M, Cater R, et al. The electronic patient record in primary care—regression or progression? A cross sectional study. *BMJ* 2003; 326: 1439–43.
29. Chauvin M. State of the art of pain treatment following ambulatory surgery. *Eur J Anaesthesiol Suppl* 2003; 28: 3–6.
30. Huang N, Cunningham F, Laurito CE, Chen C. Can we do better with postoperative pain management? *Am J Surg* 2001; 182: 440–8.
31. Shaikh S, Chung F, Imaengiyae C, Yung D, Bernstein M. Pain, nausea, vomiting and ocular complications delay discharge following ambulatory microdiscectomy. *Can J Anesth* 2003; 50: 514–8.
32. Janssen KJ, Kalkman CJ, Grobbee DE, Bonsel GJ, Moons KG, Vergouwe Y. The risk of severe postoperative pain: modification and validation of a clinical prediction rule. *Anesth Analg* 2008; 107: 1330–9; Lippincott Williams & Wilkins.
33. 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.
34. 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.
35. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; 247: 2543–6.
36. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.
37. 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.
38. 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.
39. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000; 19: 3401–15.