Prognosis and prognostic research: validating a prognostic model

Douglas G Altman,[1] Yvonne Vergouwe,[2] Patrick Royston,[3] Karel G M Moons[2]

Prognostic models are of little clinical value unless they are shown to work in other samples. **Douglas Altman and colleagues** describe how to validate models and discuss some of the problems

[1]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD
[2]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, Netherlands
[3]MRC Clinical Trials Unit, London NW1 2DA
Correspondence to: D G Altman doug.altman@csm.ox.ac.uk
Accepted: 6 October 2008

Cite this as: *BMJ* 2009;338:b605
doi: 10.1136/bmj.b605

This article is the third in a series of four aiming to provide an accessible overview of the principles and methods of prognostic research

Prognostic models, like the one we developed in the previous article in this series,[1] yield scores to enable the prediction of the risk of future events in individual patients or groups and the stratification of patients by these risks.[2] A good model may allow the reasonably reliable classification of patients into risk groups with different prognoses. To show that a prognostic model is valuable, however, it is not sufficient to show that it successfully predicts outcome in the initial development data. We need evidence that the model performs well for other groups of patients.[1 3] In this article, we discuss how to evaluate the performance of a prognostic model in new data.[4 5]

Why prognostic models may not predict well

Various statistical or clinical factors may lead a prognostic model to perform poorly when applied to other patients.[4 6] The model's predictions may not be reproducible because of deficiencies in the design or modelling methods used in the study to derive the model, if the model was overfitted, or if an important predictor is absent from the model (which may be hard to know).[1] Poor performance in new patients can also arise from differences between the setting of patients in the new and derivation samples, including differences in healthcare systems, methods of measurement, and patient characteristics. We consider those issues in the final article in the series.[7]

Design of a validation study

The main ways to assess or validate the performance of a prognostic model on a new dataset are to compare observed and predicted event rates for groups of patients (calibration) and to quantify the model's ability to distinguish between patients who do or do not experience the event of interest (discrimination).[8 9] A model's performance can be assessed using new data from the same source as the derivation sample, but a true evaluation of generalisability (also called transportability) requires evaluation on data from elsewhere. We consider in turn three increasingly stringent validation strategies.[4]

Internal validation—A common approach is to split the dataset randomly into two parts (often 2:1), develop the model using the first portion (often called the "training" set), and assess its predictive accuracy on the second portion. This approach will tend to give optimistic results because the two datasets are very similar. Non-random splitting (for example, by centre) may be preferable as it reduces the similarity of the two sets of patients.[1 4] If the available data are limited, the model can be developed on the whole dataset and techniques of data re-use, such as cross validation and bootstrapping, applied to assess performance.[1] Internal validation is helpful, but it cannot provide information about the model's performance elsewhere.

Temporal validation—An alternative is to evaluate the performance of a model on subsequent patients from the same centre(s).[6 10] Temporal validation is no different in principle from splitting a single dataset by time. There will clearly be many similarities between the two sets of patients and between the clinical and laboratory techniques used in evaluating them. However, temporal validation is a prospective evaluation of a model, independent of the original data and development process. Temporal validation can be considered external in time and thus intermediate between internal validation and external validation.

External validation—Neither internal nor temporal validation examines the generalisability of the model, for which it is necessary to use new data collected from an appropriate (similar) patient population in a different centre. The data can be retrospective data and so external validation is possible for prediction models that need long follow-up to gather enough outcome events. Clearly, the second dataset must include data on all the variables in the model. Fundamental design issues for external validation, such as sample selection and sample size, have received limited attention.[11]

Comparing predictions with observations

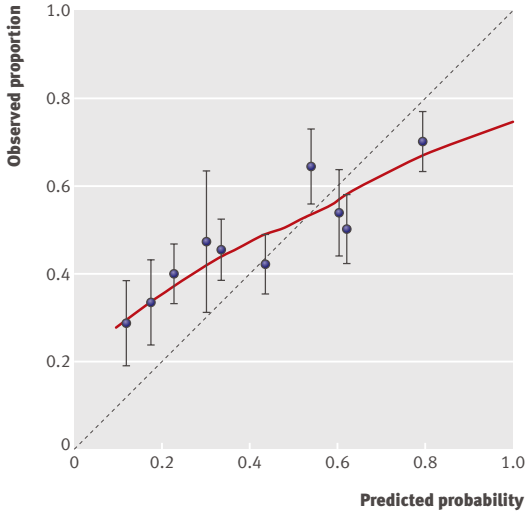
Proper validation requires that we use the fully specified existing prognostic model (that is, both the selected variables and their coefficients) to predict

outcomes for the patients in the second dataset and then compare these predictions with the patients' actual outcomes. This analysis uses each individual's event probability calculated from their risk score from the first model.[1]

Both calibration and discrimination should be evaluated.[1] Calibration can be assessed by plotting the observed proportions of events against the predicted probabilities for groups defined by ranges of predicted risk, as discussed in the previous article.[1] This plot can be accompanied by the Hosmer-Lemeshow test,[12] although the test has limited statistical power to assess poor calibration and is oversensitive for very large samples. For grouped data, as in the examples below, a χ^2 test can be used to compare observed and predicted numbers of events. It may also be helpful to compare observed and predicted outcomes in groups defined by key patient variables, such as diagnostic or demographic subgroups. Discrimination may be summarised by the *c* index (area under the receiver-operator curve) or R^2.[1]

The figure shows a typical example of a poorly calibrated model.[13] The line fitting the data is very different from the diagonal line representing perfect calibration. A slope much smaller than 1 indicates that the range of observed risks is much smaller than the range of predicted risks. The poor discriminative ability of the model was shown by a low *c* index of 0.63 (95% confidence interval 0.60 to 0.66) in the validation sample compared with 0.75 (0.71 to 0.79) in the development sample.[13]

It may be helpful to prespecify acceptable performance of a model in terms of calibration and discrimination. If this performance is achieved, the model may be suitable for clinical use. It is, however, unclear how to determine what is acceptable, especially as prognostic assessments will still be necessary and even moderately performing models are likely to do better than clinicians' own assessments.[14 15]



Calibration plot for a scoring system for predicting postoperative nausea and vomiting.[13] Circles indicate the observed frequency of events per tenth of predicted risk, with vertical lines representing 95% confidence intervals. The solid line shows the relation between observed outcomes and predicted risks

Table 1 | Predicted and observed mortality by EuroSCORE risk level for Australian patients having coronary artery bypass grafting[17]

EuroSCORE	No of deaths/ patients	Observed mortality (%) (95% CI)	Predicted mortality (%) (95% CI)
0-2 (low risk)	8/1955	0.41 (0.18 to 0.80)	1.03 (0.99 to 1.06)
3-5 (medium risk)	17/1996	0.85 (0.50 to 1.36)	3.90 (3.87 to 3.94)
≥6 (high risk)	87/1641	5.30 (4.27 to 6.50)	8.52 (8.39 to 8.65)
Total	112/5592	2.00 (1.65 to 2.40)	4.25 (4.16 to 4.34)

Case studies

We illustrate the above ideas with four case studies with various performance characteristics.

Predicting operative mortality of patients having cardiac surgery

The European system for cardiac operative risk evaluation (EuroSCORE) was developed using data from eight European countries to predict operative mortality of patients having cardiac surgery.[16] The score combines nine patient factors and eight cardiac factors; it has been successfully validated in other European cohorts. Yap and colleagues examined the performance of EuroSCORE in an Australian cohort that was different from the derivation cohort, with a generally higher risk of death.[17] For example, 41% of the Australian cohort were aged over 70 compared to 27% in the European cohort, and there were 15% *v* 10% with recent myocardial infarction. Yet the observed mortality in the Australian cohort was consistently much lower than that predicted by the EuroSCORE model (table 1). Observed mortality for three risk groups was only half the predicted mortality. The calibration of the model in these new patients was thus poor, although it retained discrimination in the new population.

There are various possible explanations for this poor performance including different epidemiology of ischaemic heart disease and differences in access to health care. Also, the EuroSCORE model was based on data from 1995 and may not reflect current cardiac surgical practice even in Europe. In such a case, however, it is easy to recalibrate the original model so that calibration and predictions become accurate in the new population, while preserving discrimination.[18 19] However, this updated model might require further validation. We will discuss this further in the next article.[7]

Predicting postoperative mortality after colorectal surgery

A prospective study recruited 1421 consecutive patients having colorectal surgery for cancer or diverticular disease from 81 centres in France in 2002.[20] A multiple logistic regression analysis on a large number of factors identified four that were significantly predictive of postoperative mortality. All were binary, although two (age

and weight) were originally continuous. The investigators found that the number of the four factors present was a strong predictor of mortality (table 2).

The model development can be criticised: four variables were selected from numerous candidates, the number of deaths was small, continuous variables were dichotomised, and the authors replaced the regression model by a simple count of factors present, neglecting the relative weights (regression coefficients) of the four predictors. Nevertheless, when this risk score was tested in a new series of 1049 patients recruited from 41 centres in 2004,[21] the mortality across the score categories (a kind of calibration) was similar to that in the original study (table 2). Both datasets show a strong risk gradient with good discrimination, but for one category the observed and predicted event probabilities are quite different. This example shows the difficulty of judging how well a model validates.

Predicting failure of non-invasive positive pressure ventilation

Non-invasive positive pressure ventilation may reduce mortality in patients with exacerbation of chronic obstructive pulmonary disease, but it fails in some patients. A prognostic model was developed to try to identify patients at high risk of failure of ventilation, both at admission and after two hours. Using data from 1033 patients admitted to 14 different units, researchers used stepwise logistic regression to develop a model comprising four continuous variables (APACHE II score, Glasgow coma scale, pH, and respiratory rate) each grouped into two or three categories.[22] The model for failure after two hours of ventilation had a *c* index of 0.88. Predicted probabilities of events varied widely from 3% to 99% for different combinations of variables.

The same researchers validated their model using data from an independent sample of 145 patients admitted to three units—it is unclear whether these were among the original 14 units. The Hosmer-Lemeshow test showed no significant difference (P>0.9) between observed and expected numbers of failures, and the *c* index of 0.83 was similar to that observed in the original sample. The high discrimination suggests that the model could help decide clinical management of patients. However, the size of their validation sample may be inadequate to support strong inferences.

Table 2 | Mortality after colorectal surgery in relation to number of risk factors present in two cohorts[20 21]

	Initial cohort		Validation cohort	
No of risk factors	No of deaths/ patients	Mortality (%)	No of deaths/ patients	Mortality (%)
0	3/580	0.5	2/424	0.5
1	11/557	2.0	6/366	1.6
2	20/223	9.0	11/153	7.2
3	9/56	16.1	22/47	46.8
4	5/10	50.0	7/10	70.0
Total	48/1426	3.3	48/1000	4.8

Table 3 | Number (percentage) of preschool children developing complications after presenting to primary care with acute cough in relation to signs at presentation

Signs present	Initial cohort	Validation cohort
Neither sign	10/153 (6)	13/95 (14)
Chest signs only	6/33 (18)	4/29 (14)
Fever only	5/18 (28)	1/11 (9)
Both signs	2/5 (40)	0/8 (0)
Total	23/209 (11)	18/143 (13)

Predicting complications of acute cough in preschool children

To reduce clinical uncertainty concerning preschool children presenting to primary care with acute cough, Hay and colleagues derived a clinical prediction rule for complications.[23] They used logistic regression to examine several potential predictors and produced a simple classification using two binary variables (fever and chest signs) to create four risk groups. Risk of complications varied from 6% with neither symptom to 40% with both (table 3). The *c* index was 0.68.

Unfortunately, evaluation of the model in a second dataset failed to confirm the value of this classification (table 3).[24] The authors suggested several explanations, including the possibility that doctors might preferentially have treated symptomatic patients with antibiotics. It may simply be that the primary data included too few children who developed complications to allow reliable modelling.

Discussion

Validation studies are necessary because performance in the original data may well be optimistic,[6] but temporal and (especially) external validation studies are scarce.[25]

It seems to be widely believed that the statistical significance of predictors in a multivariable model shows the usefulness of a prediction model. Also, when evaluating a model with new data authors seem to want to calculate P values and conclude that the validation is satisfactory if there is no significant difference between, say, observed and predicted event rates, for example based on the Hosmer-Lemeshow test. Neither view is correct—P values do not provide a satisfactory answer.

Rather, in a validation study we evaluate whether the performance of the model on the new data (its calibration and, especially, discrimination) matches, or comes close to, the performance in the data on which it was developed. But even if the performance is less good, the model may still be clinically useful.[4] The assessment of usefulness of a model thus requires clinical judgment and depends on context.

A model is "a snapshot in place and time, not fundamental truth."[26] If the case mix in the validation sample differs greatly from that of the derivation sample the model may fail, although it may be possible to improve the model by simple recalibration, as in the EuroSCORE example above, or even by including new variable(s) that relate to the different case mix and are found to be prognostic in the new

SUMMARY POINTS

Unvalidated models should not be used in clinical practice

When validating a prognostic model, calibration and discrimination should be evaluated

Validation should be done on a different data from that used to develop the model, preferably from patients in other centres

Models may not perform well in practice because of deficiencies in the development methods or because the new sample is too different from the original

sample.[27] For example, the range of patients' ages in the derivation and validation samples might differ markedly, so that age might not be recognised in the derivation set as an important prognostic factor. In addition, performance of a model may change over time and re-evaluation may be indicated after some years. We consider these possibilities further in the next article.[7]

Simplicity of models and reliability of measurements are important criteria in developing clinically useful prognostic models.[2 28] Experience shows that more complex models tend to give overoptimistic predictions, especially when extensive variable selection has been performed,[29] but there are notable exceptions.

As the aim of most prognostic studies is to create clinically valuable risk scores or indexes, the definition of risk groups should ideally be driven mainly by clinical rather than statistical criteria. If a clinician would leave untreated a patient with at least a 90% chance of surviving five years, would apply aggressive therapy if the prognosis was 30% survival or less, and would use standard therapy in intermediate cases, then three prognostic groups seem sensible. Validation of the model would investigate whether the observed proportions of events were similar in groups of patients from other settings and whether separation in outcome across those groups was maintained.

Few prognostic models are routinely used in clinical practice, probably because most have not been externally validated.[25 28] To be considered useful, a risk score should be clinically credible, accurate (well calibrated with good discriminative ability), have generality (be externally validated), and, ideally, be shown to be clinically effective—that is, provide useful additional information to clinicians that improves therapeutic decision making and thus patient outcome.[25 28] It is crucial to quantify the performance of a prognostic model on a new series of patients, ideally in a different location, before applying the model in daily practice to guide patient care. Although still rare, temporal and external validation studies do seem to be becoming more common.

DGA is supported by Cancer Research UK. KGMM and YV are supported by the Netherlands Organization for Scientific Research (ZON-MW 917.46.360). PR is supported by the UK Medical Research Council. We thank Yves Panis and Alastair Hay for clarifying some details of the case studies.

Contributors: The articles in the series were conceived and planned by DGA, KGMM, PR and YV. DGA wrote the first draft of this paper. All the authors contributed to subsequent revisions. DGA is the guarantor.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604.
2 Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why and how? *BMJ* 2009;338:b375.
3 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.
4 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453-73.
5 Altman DG, Royston P. Evaluating the performance of prognostic models. In: Rothwell P, ed. *Treating individuals: from randomised trials and systematic reviews to personalised medicine in routine practice*. *Lancet*: Edinburgh, 2007:213-29.
6 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24.
7 Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prediction models in clinical practice. *BMJ* 2009;338:b606.
8 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.
9 Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. *J Clin Epidemiol* 1997;50:21-9.
10 Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991;10:1213-26.
11 Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-83.
12 Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. New York: Wiley, 2000.
13 Van den Bosch JE, Kalkman CJ, Vergouwe Y, Van Klei WA, Bonsel GJ, Grobbee DE, et al. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. *Anaesthesia* 2005;60:323-31.
14 Ivanov J, Borger MA, David TE, Cohen G, Walton N, Naylor CD. Predictive accuracy study: comparing a statistical model to clinicians' estimates of outcomes after coronary bypass surgery. *Ann Thorac Surg* 2000;70:162-8.
15 Loeb M, Walter SD, McGeer A, Simor AE, McArthur MA, Norman G. A comparison of model-building strategies for lower respiratory tract infection in long-term care. *J Clin Epidemiol* 1999;52:1239-48.
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;16:9-13.
17 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;29:441-6.
18 Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 1995;14:1999-2008.
19 Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567-86.
20 Alves A, Panis Y, Mathieu P, Mantion G, Kwiatkowski F, Slim K. Postoperative mortality and morbidity in French patients undergoing colorectal surgery: results of a prospective multicenter study. *Arch Surg* 2005;140:278-83.
21 Alves A, Panis Y, Mantion G, Slim K, Kwiatkowski F, Vicaut E. The AFC score: validation of a 4-item predicting score of postoperative mortality after colorectal resection for cancer or diverticulitis: results of a prospective multicenter study in 1049 patients. *Ann Surg* 2007;246:91-6.
22 Confalonieri M, Garuti G, Cattaruzza MS, Osborn JF, Antonelli M, Conti G, et al. A chart of failure risk for noninvasive ventilation in patients with COPD exacerbation. *Eur Respir J* 2005;25:348-55.
23 Hay AD, Fahey T, Peters TJ, Wilson A. Predicting complications from acute cough in pre-school children in primary care: a prospective cohort study. *Br J Gen Pract* 2004;54:9-14.
24 Hay AD, Gorst C, Montgomery A, Peters TJ, Fahey T. Validation of a clinical rule to predict complications of acute cough in preschool children: a prospective study in primary care. *Br J Gen Pract* 2007;57:530-7.
25 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.
26 Iezzoni LI. Statistically derived predictive models. Caveat emptor. *J Gen Intern Med* 1999;14:388-9.
27 Hubacek J, Galbraith PD, Gao M, Humphries K, Graham MM, Knudtson ML, et al. External validation of a percutaneous coronary intervention mortality prediction model in patients with acute coronary syndromes. *Am Heart J* 2006;151:308-15.
28 Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;311:1539-41.
29 Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Appl Stat* 1999;48:313-29.