

What do we mean by validating a prognostic model?

Douglas G. Altman^{1,*†} and Patrick Royston²

¹*ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, U.K.*

²*Department of Medical Statistics and Evaluation, Imperial College School of Medicine, Hammersmith Hospital, Ducane Road, London W12 0NN, U.K.*

SUMMARY

Prognostic models are used in medicine for investigating patient outcome in relation to patient and disease characteristics. Such models do not always work well in practice, so it is widely recommended that they need to be validated. The idea of validating a prognostic model is generally taken to mean establishing that it works satisfactorily for patients other than those from whose data it was derived. In this paper we examine what is meant by validation and review why it is necessary. We consider how to validate a model and suggest that it is desirable to consider two rather different aspects – statistical and clinical validity – and examine some general approaches to validation. We illustrate the issues using several case studies. Copyright © 2000 John Wiley & Sons, Ltd.

‘Validation is one of those words... that is constantly used and seldom defined.’
ALVAN FEINSTEIN [1]

1. INTRODUCTION

Regression analysis is much used to develop models for prediction of outcome from one or more explanatory variables. There is a huge literature on both methodology and applications. Indeed, the use of regression models, including logistic and Cox (proportional hazards) models, is now widely seen as a routine statistical analysis. Concerns have been raised about both the overuse and misuse of this technology [2–4].

In medicine, regression models relating to patient outcome are termed prognostic models. They are widely used in cancer and other medical specialties for investigating patient outcome in relation to patient and disease characteristics. Prognostic models may be developed for scientific or clinical reasons, or both. In some studies the aim is to gain insight into the disease process by determining which variables are associated with prognosis, or to determine whether a particular variable is prognostic after allowance for other, previously identified prognostic variables. The

* Correspondence to: Douglas G. Altman, ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, U.K.

† E-mail: altman@icrf.icnet.uk

primary focus of this paper is models which are developed to predict the outcome of future patients. Reasons for wishing to make such predictions include:

- (i) to inform treatment or other clinical decisions for individual patients;
- (ii) to inform patients and their families;
- (iii) to create clinical risk groups for informing treatment or for stratifying patients by disease severity in clinical trials.

Just as one can fit a straight line through any bivariate data, so one can derive a multiple regression model from any data set. It is clear that such a model will have no clinical value unless it has been shown to predict outcome with some success. As Burstein [5] noted, 'Any classification system, be it nominal, ordinal, or scalar, should be proved to be a workable tool before it is used in a discriminatory or predictive manner.' Usefulness is determined by how well a model works in practice, not by how many zeros there are in the associated *P*-values.

There are two broad ways in which a model may be useful. First, it may allow the reliable classification of patients into two or more groups with different prognoses. Such classification schemes can be used to influence therapy or save patients from unnecessary referrals or tests. Second, a prognostic model can be used to estimate the prognosis of individual patients. While in some sense these are two different ways of looking at the same information, they differ fundamentally.

A basic issue relating to predicting events is that binary outcome data have considerable irreducible variability. The data are all zeros and ones, but the predictions are probabilities lying between these extremes. We thus need to distinguish between how well a model may predict for groups of patients, and how well for individual patients. Suppose that we have a prognostic model which predicts that, say, 69 per cent of a defined set of patients will survive five years from diagnosis of a certain cancer. We would predict that among a large group of similar patients, close to 69 per cent will indeed still be alive after five years. We can increase the precision by increasing the sample size so that there may be little uncertainty in the prediction, but what can we say about an individual patient? Even a 'valid' model can tell us, and the patient, only that there is a 69 per cent chance of surviving five years. This does not provide much certainty about whether or not they will survive that long. Thus although with an excellent model we may successfully distinguish between high and low risk patients, and can estimate group survival probabilities with precision, our ability to provide informative prognoses at the individual level, such as the patient's expected survival time with a narrow 95 per cent confidence interval, is almost always limited. We believe that the distinction between what is achievable at the group and individual levels is not well understood.

A vital aspect of prediction is to consider whether a model derived from an analysis of the original data set is transportable to similar patients in another location. The concept is sometimes referred to as generalizability or validity, and a model which is found to pass such a test is said to have been validated. The main purposes of this paper are to consider: (i) what is (or should be) meant by validating a model; (ii) why we need to validate models, and (iii) how we should attempt to validate a model. We illustrate these issues using several case studies.

2. WHAT DO WE MEAN BY VALIDATION?

The idea of validating a prognostic or diagnostic model is generally taken to mean establishing that it works satisfactorily for patients other than those from whose data the model was derived.

There are several widely used approaches to assessing the performance (prediction accuracy) of a prognostic model. They include comparison of observed and predicted event rates for groups of patients (calibration) or individuals (accuracy scores), and measures which distinguish between patients who do or do not experience the event of interest (discrimination) [6, 7]. It is not our intention to review in detail the various statistical methods that can be used.

There is a difference between the process of determining whether a model is indeed valuable in predicting outcome and the judgement that it actually is so. We think it is misleading to say that a model was validated, if by this it is meant simply that its performance was evaluated. Another semantic issue is the contrast between the term validation used in the present context and the psychometric concept of validity, widely used in studies of measurement. Validity is the property of a measurement method to measure what it is intended to measure [8, 9]. It is customary to use a correlational approach, effectively judging within-subject variation against between-subject variation. In model fitting, especially in medicine, such an approach is not adequate, as the issue of primary importance is the quality of predictions for individuals or subgroups.

In an ideal world, we might imagine creating a model with data from a well-designed prognostic factors study with adequate numbers of patients. We might hope that the model encapsulated all the prognostic information in the data and ignored all the 'noise' variables. Specifically, we might think we had all the right variables in the model (perhaps including appropriate interactions between terms) and that we had determined correctly the functional form of the effect of every continuous variable on the outcome. Furthermore we might hope that the model could not be improved by measuring any other factors and including them in the model. With this correct model, we would now expect to predict the outcome for future patients with as much precision as the prognostic information inherent in our measurements would allow.

Validation of such a model, with the belief of its 'correctness', would then involve the collection of an appropriate sample of new patients, the measurement of all the required prognostic variables (using the same techniques, such as laboratory assays or clinical assessments, as before) and the comparison of predictions with outcomes using one or more of the ways indicated above. We consider how to validate a model in Section 4.

2.1. Valuable, or merely valid?

In practice, a validated model as just defined may be of no practical worth. If the intrinsic prognostic information is weak, the predictions, even if unbiased, will not enable patients to be separated into clinically useful prognostic groups. By contrast, if the prognostic information is strong, even a biased model may yield clinically useful separation. The development of a successful model depends on the following features: the potential for accurate prognosis, which is presumably unknown; the intrinsic prognostic information in the available factors, which depends on many things, including the physiology of the disease in question; the measurement process, which converts the intrinsic information into numbers, some measurements being more reliable than others; and the accuracy with which the model converts the measurements into predictions. Two related questions arise:

1. With the available factors, is the model the best that can be found?
2. Does the model predict accurately enough for its purpose?

A useful definition of validity depends crucially on one's view of the aims of a prognostic model. Question 1 implies a 'statistical' aim – a model which failed to satisfy certain statistical notions of

correctness would be deemed invalid. Question 2 implies a clinical aim – a model which failed to be useful in a clinical setting would be regarded as invalid. A model might fail either because it is statistically invalid (which we can usually do something about), or because the intrinsic prognostic information is weak (which we cannot). Furthermore, the same model might fail according to one clinical criterion and pass according to another. These considerations lead us to define two types of validated model.

2.2. *Two types of validated model*

The foregoing analysis implies that the judgement of validity or otherwise of a model depends on circumstances. For example, the discrepancy between relapse rates of 90 per cent (predicted) and 75 per cent (observed) may be acceptable if the aim is to identify sufficiently separated prognostic subgroups, and such separation is seen. However, if the aim is to identify patients with at least an 85 per cent chance of relapse, the model would not be considered valid for its purpose. We see the task of the statistician within this paradigm as twofold – first, to construct a prognostic model and estimate the amount of intrinsic prognostic information on a clinically meaningful scale, and second, to provide measures of agreement between observations and predictions appropriate to particular categories of clinical aim.

The first task involves the construction of accurate prognostic models, with an attempt to allow for the overestimation of prognostic information – the well known problem of overoptimistic prediction when a model is fitted and evaluated on the same data set. The second task involves the classification of types of clinical aim and the provision of appropriate summary statistics which the clinician can use to judge the validity of the model. We return to these issues in Section 4.

We may now propose definitions of two types of validated model:

1. A *statistically validated model* is one which passes all appropriate statistical checks, including goodness-of-fit on the original data set and unbiased prediction on a new data set.
2. A *clinically validated model* is one which performs satisfactorily on a new data set according to context-dependent statistical criteria laid down for it.

Note that with the above definitions, a clinically validated model may be statistically invalid (for example, its predictions are biased, or it fails a goodness-of-fit test) and a statistically validated model may be clinically invalid (for example, the intrinsic prognostic information is too weak). Generally it may be harder to derive a statistically validated model than a clinically validated one, one reason being the considerable difficulty of overcoming the problem of overoptimism and bias at the model-building stage. However, a clinically validated model is likely to be more useful than a statistically validated one.

3. WHY DO WE NEED TO VALIDATE A PROGNOSTIC MODEL?

There are several reasons why the performance of a prognostic model needs to be evaluated before its results can be used. The general point is that we need evidence that it does indeed do what it is intended to – that its predictions are adequate for the purpose (as discussed below). There are several inter-related reasons why prognostic models may not perform well.

3.1. Deficiencies of standard modelling methods

It is known that analyses which are not prespecified but are data-dependent are liable to lead to overoptimistic conclusions. Standard statistical methods used to derive prognostic models all have data-dependent aspects, so we would expect them to give an overoptimistic assessment of predictive performance. There is also a wealth of empirical evidence to back this up. We discuss some relevant case studies later in this paper.

For clinical purposes it is usually necessary for a predictive model to be based on a small number of variables, and it is arguable that parsimony is a desirable feature of a good model [10]. In most cases, however, there is a large number of 'candidate' variables available for consideration, and thus there is a need to select the 'important' ones. The data-dependent aspect of most models stems from this variable selection.

Of the various strategies for producing a prognostic model, by far the most common is multiple regression using a stepwise selection algorithm (backward or forward). The choice of variables included in the final model is based on multiple sequential hypothesis testing of individual variables, usually with $P < 0.05$ as the inclusion criterion. This is a fully automated procedure requiring no intellectual input. There is no reason why it should yield the model which is best in a predictive sense – indeed the procedure is best only regarding its convenience. It is possible, and also desirable, to use clinical criteria or statistical methods to reduce the number of candidate variables, thus reducing the risk of an overoptimistic model [11].

An alternative approach is all subsets regression, although it is rarely used (an example is Gamel *et al.* [12]). This method can discover combinations of variables that explain more variation in patient outcome than the model achieved by stepwise selection. It is also a fully automated procedure once one has defined the criterion for choosing the best model, and uses data-dependent variable selection. A popular approach is to choose the model which optimizes a measure of goodness-of-fit penalized for including each extra variable, such as the Akaike information criterion or Mallows' C_p . However, the approach has some major drawbacks, including the possibility of selecting models which omit important predictors [13].

Recent alternative methodologies include regression trees (CART) and neural networks. These methods also use the data to determine the model, although in rather different ways. Their increased use in recent times may be more due to their novelty than to evidence that they produce 'better' models. The evidence so far suggests that they do not offer any consistent advantage in this context [14–17].

3.2. Deficiencies in the design of prognostic studies

The most reliable observational studies are those that attempt to emulate the careful design standards used in clinical trials, with the goal of achieving the same answer as if an experimental study had been performed [18]. Simon and Altman [19] highlighted various weaknesses of prognostic factor studies which could result in misleading findings – creating overoptimism and/or bias. These include the absence of clear inclusion and exclusion criteria, many patients excluded through missing data (which may not be missing at random), unclear rationale for the choice of treatments, and inadequate sample size. The definition of the characteristics of the sample is of clear importance to the clinician who wishes to know whether a model is relevant to a particular patient.

The problems that arise from data-dependent selection are exacerbated by small sample size. With a small sample there will be a low signal-to-noise ratio, with an increased risk of selecting

unimportant variables and failing to include important ones. Another consideration is the number of events (for example, deaths) per variable (EPV) considered for inclusion in the model. Harrell *et al.* [11] concluded that for regression modelling the EPV should be at least ten times the number of potential prognostic variables that could be included in the model. Peduzzi *et al.* [20, 21] reported simulation studies that showed that parameter estimates in proportional hazards and logistic regression models are unreliable when $EPV < 10$. They thus also suggested a minimum of $EPV = 10$. Feinstein [1] suggested that an EPV of 20 is safer. Most published studies do not meet either criterion.

3.3. Models may not be transportable

Even when the methodology of a study is impeccable, there are other reasons why a model may not be transportable to other locations. Foremost here is the degree of dissimilarity between the patients in different centres, otherwise known as variation in 'case-mix'. Clearly, if the prognostic model includes all important prognostic variables (and has modelled them appropriately) then the model should be transportable to centres with a different case-mix. If, however, one or more important variables is not present in the model then variation in case-mix could lead to quite different model performance in different centres. The difficulty, of course, is that one can never know whether a model does indeed include all important variables.

Some steps can be taken to reduce the data-dependence of a model (as discussed below). However, possible weaknesses in the design and analysis plus the risk of omitted variables mean that it is not possible to use the original data to determine reliably the transportability of a model. These considerations argue strongly for the need to evaluate performance of a model on a new series of patients, ideally in a different location. In addition, for a model to be adopted by others requires a degree of confidence in its reliability. Such credibility can probably only be gained by empirical demonstration of transportability [22].

4. HOW SHOULD WE VALIDATE A MODEL?

There are several main considerations in validating a model:

- (i) study design;
- (ii) measuring the intrinsic prognostic information;
- (iii) comparing predictions with observations;
- (iv) quantifying the performance of a model;
- (v) prespecifying adequate performance.

We do not intend in this paper to delve deeply into the technical aspects of the above issues. However, to provide a focus for the case studies presented in the next section, we believe it is helpful to air them briefly. We shall concentrate on the second consideration, and within that, on the simplest statistical measures of prognostic information. We shall avoid the question of how 'best' to construct a classification scheme, assuming instead that prognostic subgroups have already been defined somehow. Similarly our approach to the comparison of predictions with observations will again be deliberately oversimplified. Graf *et al.* [23] discuss the issue at some length in the context of survival models.

4.1. Study design

We will consider a hierarchy of increasingly stringent validation strategies:

- (i) internal – procedures restricted to a single data set;
- (ii) temporal – evaluation on a second data set from the same centre(s);
- (iii) external – evaluation on data from one or more other centres, perhaps by different investigators.

4.1.1. Internal validation. One common way of establishing how well a model might perform for further patients is data splitting or cross-validation. Here the original sample is split into two parts before the modelling begins. The model is derived on the first portion of the data (often called the ‘training’ set) and then its ability to predict outcome is evaluated on the second or ‘test’ data set. A variation is to carry out the modelling procedure on each portion of the data and to evaluate each model on the other portion. An issue is how to split the data set. Although cross-validation is widely recommended, authors rarely consider what proportion of patients should be in the test and training sets (or fail to justify any recommendation), see Cox [24]. Random splitting must lead to data sets that are the same other than for chance variation and is thus a weak procedure [1, 25]. Furthermore, estimates of predictive accuracy from data-splitting procedures, though unbiased, tend to be imprecise (see Efron and Tibshirani [26]). A tougher test is to split the data in a non-random way. For example, we might take groups of patients seen in different time periods. Rather different, and better, approaches are to use bootstrapping or ‘leave-one out’ cross-validation. From these analyses shrinkage factors can be estimated and applied to the regression coefficients to counter overoptimism [27, 28]. These techniques allow evaluation on multiple data sets, but are still an internal procedure.

4.1.2. Temporal validation. An alternative is to evaluate the performance of a model on subsequent patients within the same centre(s) [29]. Unfortunately, it is no different in principle from the idea already mentioned of splitting a single data set into two cohorts seen in different time periods. A better model would be obtained simply by analysing all the available data because the sample size would be larger. However, it is at least a prospective evaluation, independent of the original data and model-fitting process. A disadvantage of the approach when the outcome is survival time is the need to wait several years to accrue an adequate number of events in a further cohort.

4.1.3. External validation. Neither internal nor temporal evaluation addresses the wider issue of the generalizability of the model. As the goal of validation is to demonstrate satisfactory performance for patients from a different population from the original, it is clearly desirable to evaluate a model on new data collected from an appropriate patient population in a different centre. Important design issues such as sample selection and sample size have been largely neglected in the literature. External evaluation can be based on retrospective data and so is viable for validating survival models needing long follow-up.

4.2. Measuring intrinsic prognostic information

For the purpose of the following discussion we shall assume that prognoses are to be framed as predicted probabilities of a particular event, implicitly or explicitly linked to a specific time-point

– for example, the chance of surviving for 5 years following initial treatment. The predicted probabilities are obtained as outputs from a prognostic model. Such a model may range from the simplest possible, for example two groups of patients defined by a single binary prognostic factor, to exotic constructions in which a prognostic index is defined as a linear or non-linear function of many predictors with shrinkage factors applied to the regression coefficients. We do not need to distinguish conceptually between models for survival time (Cox regression) and binary regression models (logistic regression), because survival models can generate predicted probabilities at any given time-point within the follow-up period of the study. Graf *et al.* [23] stated that attempts to predict survival times for individual patients are inadequate. There may be exceptions in studies with long follow-up times and a suitable parametric survival model to provide useful individual predictions. This is not the normal situation and we do not consider such predictions here.

Intuitively, the idea of prognostic information is straightforward and relates to the spread of predicted probabilities. For example, in an analysis unadjusted for other factors, the estimated chance of surviving for 3 years following initial treatment for node-positive breast cancer may be about 90 per cent for patients with 1–3 affected lymph nodes compared with about 60 per cent for those with 10 or more affected nodes. By contrast, the corresponding figures for pre- and post-menopausal patients may be about 84 per cent and 82 per cent, respectively. The prognostic information contained in lymph node status is clearly much greater than that in menopausal status, since the spread of probabilities is 0.3 as against only 0.02.

However, many difficulties arise with such a naïve approach. The spread of probabilities depends on how finely the prognostic factor or index is graded: the finer, the greater the spread. In the limit, the spread could be as wide as that of all the patient-specific predictions in the study. It also depends on the prevalence of the event. In a survival study, the spread usually increases with the length of follow-up. Equally crucially, it is affected in a potentially major way by the amount of overoptimism in the statistical model used to estimate the probabilities. The overoptimism itself is influenced by many factors, including the simplicity or otherwise of the final prognostic model, the sample size or number of events, the model building strategy used (in particular, how many data-driven decisions were taken and the nominal P value, if stepwise variable selection was used), the study design, and other factors. See Sauerbrei [10] for a detailed discussion of these issues in the context of model building in survival analysis.

Nevertheless, the idea of greater or lesser separation between prognostic groups as a measure of prognostic information remains attractive, being both interpretable and pragmatic. We acknowledge the concomitant difficulties and accept that the ‘best’ definition of prognostic information remains an open research question. However, for illustration later and to further the discussion of validation we shall define a simple index of separation, PSEP. We emphasize that PSEP is not intended to be a definitive proposal.

Suppose the outcome of concern is death within a predefined period following measurement of prognostic factors. Suppose that a prognostic classification scheme of some kind has been defined in some way – directly from individual predictors, by way of a multi-factorial prognostic index or some other procedure such as CART, or the expert opinion of a physician. All we require is that any patient may be classified into one of two or more prognostic groups, and that the groups with best and worst predicted prognosis have been identified. Let

p_{worst} = predicted probability of dying for a patient in the group with the worst prognosis

p_{best} = predicted probability of dying for a patient in the group with the best prognosis.

Then the predicted prognostic information can be measured by the separation PSEP

$$\text{PSEP} = p_{\text{worst}} - p_{\text{best}}.$$

With just two groups, p_{worst} and p_{best} are closely related to the familiar concepts of the positive (PPV) and negative (NPV) predictive value of a diagnostic test by the relations $p_{\text{worst}} = \text{PPV}$ and $p_{\text{best}} = 1 - \text{NPV}$. Thus $\text{PSEP} = \text{PPV} + \text{NPV} - 1$. Given the overall prevalence of events, PPV and NPV and hence PSEP may be calculated from the sensitivity and specificity by standard formulae.

4.3. Comparing predictions with observations

With the above definitions, evaluation consists of comparing the appropriate observed and predicted measure, an aspect of model calibration. As already mentioned, the aim of forming a prognostic model is critical in considering whether it is validated. In some studies, the aim may be to identify just one important prognostic subgroup. For example, in the first case study of the next section the authors wanted to identify a small subset of patients certain to die within a short period (that is, those with $p_{\text{worst}} = 100$ per cent). In this extreme situation, a failure to validate the original model would occur if any patient with $p_{\text{worst}} = 100$ per cent in the validation sample survived. A more typical case is the attempt to discriminate between patients with good or poor prognosis, either to provide scientific information or to guide treatment decisions and/or management of the patient. Here the scope is wider and the question of whether or not the model has been successfully validated is trickier to answer.

As a hypothetical example, suppose three prognostic groups with predicted three-year survival probabilities of 90, 60 and 30 per cent have been identified in a small initial study 'to develop a prognostic index for death from disease X'. Thus $p_{\text{worst}} = 0.7$, $p_{\text{best}} = 0.1$ and $\text{PSEP} = 0.6$. Suppose that a different investigator conducts an adequately designed validation study and finds corresponding three-year survival probabilities of 70, 60 and 50 per cent, all significantly different from one another. PSEP is now only 0.2, so there was presumably considerable overoptimism in the original estimate of prognostic separation. Nevertheless there is evidence that the prognostic index 'works', at least in terms of indicating differential survival rates in the anticipated order. However, if patients with only a 30 per cent chance of three-year survival could justifiably be given aggressive therapy with dangerous and unpleasant side-effects, whereas such a decision was unjustified for those with a 50 per cent chance, the prognostic model would not be validated as a tool to assist in defining the treatment decision. Of course, if the respective probabilities were, say, 65, 68 and 62 per cent we would presumably always conclude that the original prognostic model was not valid, no matter what the purpose of the model; the probabilities are approximately equal and not in the anticipated order.

4.4. Quantifying the performance of a model

It seems to us that if the principle is accepted that validation cannot be determined by statistical criteria alone but must be considered in relation to the clinical aims, what remains are statistical technicalities. This is not to say, of course, that they are trivial. For example, the comparison between predicted and observed probabilities may be made in several ways, such as at the patient level by using the Brier score [30, 23], a quadratic loss function defined as the mean squared difference between observed patient outcomes in the validation sample and the corresponding

probabilities predicted by the model. See Efron [31] for a description of other measures of explained variation for models of binary data. The Brier score has several pleasant mathematical properties, but it has the drawback that it lacks an obvious interpretation other than in general terms – the bigger the score, the worse the quality of the prediction. A cruder but more interpretable statistic is the difference between observed and predicted probabilities at the group level (PSEP), though of course more than one measure may be used.

Perhaps the most challenging technical issues concern validation only indirectly – such as, how to produce a useful prognostic model from the original data with as little overoptimism as practicable, or how to ‘deflate’ a published estimate of prognostic separation to allow for overoptimism. We shall not pursue such issues here.

4.5. Prespecifying adequate performance

Prognostic studies may fall into two categories: pragmatic and explanatory [32]. Pragmatic studies are driven by explicit clinical aims. The idea is to prejudge the quality of predictions from a prognostic model that may or may not be acceptable. This is the notion of a clear, quantitative aim, guided by statistical principles, and is reminiscent of predefining the desired size of a treatment effect or treatment difference in a clinical trial. For example, if the aim was to identify patients with a three-year survival rate of 80 per cent, a validation study showing that the chance was actually 60 per cent would cause the model to be rejected for that purpose, even though strong prognostic information might be present.

Explanatory studies are mainly concerned with scientific understanding and hypothesis generation, to answer such questions as ‘What factors are important to predict the course of disease X? Can we discriminate reproducibly between good and bad prognosis for disease X?’. Here there is no pragmatic aim, so in a validation study we would want to examine general qualitative and quantitative aspects such as:

- (i) Are the same variables still important?
- (ii) Is the functional form of the prognostic model correct?
- (iii) Are the estimated regression coefficients compatible?
- (iv) How well does the model fit the new data?
- (v) Is the correct ordering of the prognostic groups preserved?
- (vi) Are the event rates between the prognostic groups significantly different?

It is hard to judge whether the same variables are important because differences in the predictor distribution between the original and validation samples will alter the precision with which the regression coefficients are estimated and whether the variables are selected in a stepwise procedure. At the same time, we would wish to compare summaries such as PSEP, p_{worst} and p_{best} for the original groups with those in the validation study. We would get estimates of over-optimism, find whether there was a need to calibrate the model to reduce prediction bias, and so on. According to this scenario, the validation study represents just one part of the ongoing scientific process of understanding disease X.

It is helpful to prespecify adequate performance of a model. However, it should be remembered that one feature of validation is to provide an unbiased estimate of the prediction error of the model. Miller [29] suggested that it is more helpful to think of validation in this light rather than as a test which the model will or will not pass. In other words, we should focus on measures which

quantify the performance of a model and accept that the final assessment requires clinical judgement and is context-dependent. Statistics alone cannot determine clinical validity.

5. CASE STUDIES

5.1. Predicting death from severe head injury

Gibson and Stephenson [33] described the derivation of an index to predict the chance of a patient dying within an unspecified (though brief) period following severe head injury. They pointed out the need for 'a practical "bedside" method of predicting one outcome (that is, death) with 100 per cent accuracy'. Although not stated in such terms, we assume that the authors' aim in this pragmatic study was to develop an index with $p_{\text{worst}} = 100$ per cent.

The prognostic index was a weighted score calculated from seven clinical observations mainly made within 12 hours of admission to an intensive care unit in Leeds, England (see Table II of Reference [33]): age; unreactive pupils; intracranial pressure; systolic blood pressure; Glasgow coma scale score; presence of other extracranial injuries, and presence of high-density lesions on CT scan. The weights were assigned apparently subjectively 'to reflect influence on mortality and to make the system reproducible'. Based on data from the initial cohort, a score of ≥ 14 was taken as predictive of death.

The score was derived retrospectively with a series of 187 patients and tested prospectively in the same centre with a series of 52 'comparable' patients. The value of p_{worst} (that is, the proportion of patients with a score of ≥ 14 who died) was 29/29 and 13/13 in the two series – 100 per cent in each case. Thus the authors' stated aim seems to have been achieved.

In evaluating this approach, Feldman *et al.* [34] applied the same prediction rule to 479 patients retrospectively and 131 prospectively. The resulting p_{worst} values were 70 per cent and 60 per cent, respectively. They concluded that 'the Leeds prediction model is not infallible'.

We may reasonably doubt whether *with the available factors, the model is the best that can be found* (see Section 2). Since no description is given of how the weights in the prognostic score were obtained, it is difficult to criticize the model statistically. We hypothesize that the weights were adjusted in data-dependent fashion to maximize the separation between the scores of the survivors in the retrospective series and those who died. However, it is clear that no procedures were adopted for reducing the overoptimism of their model. There is therefore reason to believe that the model is statistically invalid.

We may also doubt whether *the model predicts accurately enough for its purpose* (see Sections 2 and 4), since the original aim of predicting death with 100 per cent accuracy was certainly not achieved in the independent validation study [34]. If the clinical consequence of incorrectly predicting death is serious, for example withdrawal of further active treatment (as mentioned by Gibson and Stephenson [33]) and subsequent death of a non-negligible proportion of patients who might otherwise have survived, then certainly the model of Gibson and Stephenson [33] may also be judged clinically invalid. Had the original aim been more modest, a different conclusion might have been drawn.

Concerning the amounts of prognostic information, we obtain the values shown in Table I. It appears that in the original study the separation PSEP was considerably overestimated (0.63 and 0.72 compared with 0.54 and 0.41). We also note that the death rate of 46 per cent in Gibson and Stephenson's study was more than twice the 20 per cent seen in Feldman *et al.*'s patients, so there

Table I. Performance of prognostic classification scheme of Gibson and Stephenson [53] for death from severe head injury. Values in table are proportions (and numbers) of patients who died. PSEP is as defined in Section 4.

Index	Gibson and Stephenson [33]				Feldman <i>et al.</i> [34]			
	Retrospective		Prospective		Retrospective		Prospective	
≥ 14	1.00	(29/29)	1.00	(13/13)	0.70	(16/23)	0.60	(6/10)
< 14	0.37	(58/158)	0.28	(11/39)	0.17	(76/456)	0.19	(23/121)
Total	0.47	(87/187)	0.46	(24/52)	0.19	(92/479)	0.22	(29/131)
PSEP	0.63		0.72		0.54		0.41	
(95 per cent CI)	(0.56 to 0.71)		(0.59 to 0.86)		(0.34 to 0.72)		(0.10 to 0.72)	

was presumably a difference in case-mix between the populations sampled. As already noted, case-mix and hence prevalence are expected to affect PSEP.

5.2. Predicting relapse in asthma

Fischl *et al.* [35] described a prognostic index to predict the chance of a patient with acute bronchial asthma relapsing or requiring hospitalization following initial treatment in the emergency room. The aim was described somewhat vaguely as ‘an attempt to define guidelines for the assessment of acute asthma’. Furthermore (last sentence of the discussion):

‘Assessment of patients with acute asthma by this multi-factorial approach should allow the physician to identify rapidly those patients who are at risk for relapse and in need of hospitalisation. On the basis of our data we suggest that patients with predictor index scores of 4 or higher (calculated before therapy) should be considered for prompt hospitalisation.’

We must assume that the authors’ hope was to develop a method with a high value of p_{worst} and, presumably, a large PSEP. However, no specific target value for p_{worst} , p_{best} or PSEP was stated, which is somewhat worrying in view of the authors’ radical recommendation regarding prompt hospitalization of patients with scores of 4 or more. The study should logically be regarded as explanatory rather than pragmatic.

The prognostic index was a simple score calculated from seven clinical observations made before initial treatment: pulse rate; respiratory rate; *pulsus paradoxus*; peak expiratory flow rate; dyspnoea; accessory-muscle use, and wheezing. These seven components were selected from an initial 14 candidate variables by stepwise discriminant analysis of the successfully treated and relapsing groups. The score was obtained by dichotomizing each of the variables in the final model at data-derived cutpoints which individually ‘maximized predictive significance’, that is, so-called ‘optimal’ cutpoints were used [36]. The score took possible values of 0, ..., 7. A score of ≥ 4 was taken as predictive of relapse or the need to hospitalise. The results are shown in Table II.

The value of PSEP for this rule was a very impressive 0.92. In evaluating the approach, Centor *et al.* [37] applied the same rule to 86 patients with the results shown in Table II. The value of

Table II. Prognostic classification of relapse or hospitalization in acute asthma. Values in table are proportions (and numbers) of patients who relapsed or were admitted to hospital.

Index	Fischl <i>et al.</i> [35]		Centor <i>et al.</i> [37]	
≥ 4	0.95	(81/85)	0.52	(11/21)
< 4	0.03	(4/120)	0.28	(18/65)
Total	0.41	(85/205)	0.34	(29/86)
PSEP	0.92		0.25	
(95 per cent CI)	(0.86 to 0.98)		(0.01 to 0.49)	

PSEP has been reduced to a modest 0.25, though with wide confidence limits. Centor *et al.* concluded that ‘these findings cast doubt on the utility of the index in our emergency room and, therefore, raise questions about its use in other settings’ [37].

Again we may doubt whether *with the available factors, the model is the best that can be found* (see Section 2). A considerable amount of data-driven model selection was performed, with no attempt to reduce overoptimism. Also we may doubt whether *the model predicts accurately enough for its purpose* (see Sections 2 and 4), since the original (qualitative) aim of providing a useful index of relapse or hospitalization was arguably not achieved in the independent validation study [37]. Even so, the fact that PSEP is close to significantly different from zero ($P = 0.06$, Fisher exact test) in the validation study suggests that the score of Fischl *et al.* [35] does contain some prognostic information, but probably insufficient to fulfil the original aim.

5.3. Predicting death from acute myocardial infarction

Woo [38] described a prognostic index to predict the chance of dying in hospital after acute myocardial infarction (MI) among a Chinese population. Woo *et al.* [39] evaluated the performance of the index in the original centre (Hong Kong) and in two additional cities (Guangzhou and Shanghai in mainland China). The intention was to provide ‘an objective guide for the assessment of patients with acute MI and stratify different grades of clinical severity’ [39]. Woo [38] commented that ‘a high index would conceivably unmask the high-risk patients who deserve more attention and a more energetic regimen, while a low index would identify a low-risk group for earlier ambulation and discharge from the coronary care unit and from the hospital’. We assume that ‘ambulation’ means encouraging the patient to become mobile again. Since the authors did not suggest what specific clinical actions should be taken as a result of identifying a patient’s disease severity grade, we view the study more as explanatory than pragmatic.

The index was derived from four continuous and seven binary factors using a non-standard approach based on univariate chi-square tests of association and linear discriminant analysis of the resulting significant predictors in a multivariable model. Seven factors were rejected as not being univariately associated with mortality. All the continuous factors (age, systolic blood pressure, heart size and blood urea concentration) were initially converted to categorical form by applying cutpoints – three categories for age and two for the other predictors. Thirteen parameters were estimated in the final model.

In the evaluation, Woo *et al.* [39] divided the patients into seven prognostic subgroups according to the values of the index (< 2 , 2–3, 4–5, 6–7, 8–9, 10–11, ≥ 12), and presented the

Table III. Validation of prognostic index in acute myocardial infarction (Woo *et al.* [39]). Values are proportions (and number of deaths/number of patients) in each subgroup.

Index	Original sample		Validation samples			
	Hong Kong 1971–1980		Hong Kong 1977–1986		Guangzhou 1977–1982 and Shanghai 1978–1986	
≥ 8	0.724	(63/87)	0.613	(49/80)	0.792	(38/48)
6–7	0.539	(55/102)	0.392	(38/97)	0.397	(27/68)
4–5	0.214	(39/182)	0.159	(20/126)	0.203	(35/172)
≤ 3	0.051	(14/273)	0.061	(8/132)	0.031	(5/163)
Total	0.266	(171/644)	0.264	(115/435)	0.233	(105/451)
PSEP	0.673		0.552		0.761	
(95 per cent CI)	(0.575 to 0.770)		(0.438 to 0.606)		(0.643 to 0.879)	

corresponding mortality rates by hospital location. They used chi-square tests to compare the death rate in the original sample with that in the validation sample (pooled across locations) for each prognostic subgroup separately. They concluded that ‘on the whole, the trend in mortality of the subsets correlated well with that predicted by the prognostic index ... $P > 0.2$, although a possible lower mortality was detected in the groups with prognostic scores between 6 and 7’. We present a simplified summary of their results in Table III.

Since the number of patients in the extreme prognostic groups was very small, we have amalgamated data to produce four subgroups instead of the authors’ seven. This will of course affect the value of PSEP. We have defined p_{worst} and p_{best} as the mortality in subgroups with index ≥ 8 and index ≤ 3 , respectively.

The results are interesting. PSEP is 0.673 in the original (Hong Kong) sample and 0.552 in the validation sample from the same location, showing a reduction in separation typical of overfitting in the original analysis. Nevertheless, the prognostic information seems strong and fairly reproducible. The combined value of PSEP for the two Chinese mainland cities is 0.761, actually greater than for the original Hong Kong sample. This seems to be mainly due to the relatively high death rate among severely ill patients (index 8 or more) in the mainland cities. Woo *et al.* [39] interpreted the reduction in death rates in the 1977–1986 Hong Kong cohort compared with 1971–1980 as indicating that the prognostic index could detect secular changes in death rates stratified by disease severity. However, the difference may be caused simply by the overoptimism in the predictive ability of the model; the overall death rates have hardly changed.

We may doubt whether *with the available factors, the model is the best that can be found* (see Section 2) since categorizing continuous predictors always results in some loss of information. However, since the model seems capable of predicting probabilities of death reliably over a very wide range – PSEP would have been even greater had we derived it from more extreme subgroups – we must conclude that it seems potentially clinically valid (and valuable).

5.4. Predicting falls in elderly patients

Oliver *et al.* [40] described the development and validation of STRATIFY (St Thomas’s Risk Assessment Tool In Falling Inpatients), a ‘simple unweighted scoring system’ to predict which

Table IV. Prevalence of significant risk factors for falling among elderly cases and controls ($n = 116$) with odds ratios [40].

Variable	Controls	Cases	Odds ratio
Agitation	8%	64%	20.9
Anti-arrhythmic drugs [†]	2%	27%	20.8
Unstable gait [†]	34%	78%	6.58
Fall as presenting complaint	20%	53%	4.64
Visual impairment	4%	14%	3.55
Frequent toileting	8%	17%	2.48
Transfer and mobility score of 3 or 4 (see text)	29%	46%	2.10

[†] Not included in final model.

elderly inpatients were at high risk of having a fall. The investigators intended such a model to be used to target preventive measures at those patients at most risk. A specific aim was to identify risk factors that could be readily identified by ward nurses as part of routine nursing assessment, so potential factors requiring special equipment or detailed medical assessment were not considered.

They first conducted a case-control study in a large teaching hospital to identify risk factors and construct a model. Phase 2 was a prospective evaluation of the model in the same hospital. Phase 3 was a further prospective evaluation in a district general hospital.

Several aspects of the study can be questioned. First, they took each fall as the unit of analysis rather than the patient. While some patient characteristics are not constant over time, the effect of this approach is largely to duplicate the data of those patients who fell more than once. These would not be typical patients and thus the estimated sensitivity and specificity could be biased. Also, the aim of the study was to allow targeting of preventive measures to patients at high risk, so that patients should have been the focus of the whole study. Unfortunately, the amount of duplication in the data was not stated.

Another issue is the choice of controls. In the first study, for each patient who fell the patient in the next bed was taken as control. Presumably, this meant that patients falling more than once might well have had the same control more than once, increasing the duplication in the data set. Some controls might have been in hospital a short time. It would have been better to choose as control a patient who had been in hospital for the same length of time as the case but who had not fallen.

In phase 1 there were 116 falls. The authors identified seven risk factors which were significant ($P < 0.05$) in univariate analysis (see Table IV). Two of these variables were not included in the final model. The nurses felt that they were less able to assess instability of gait than calculate the transfer and mobility score. Also, it was felt that some nurses might not be able to tell which drugs were classified as anti-arrhythmics. The decision to take values of 3 or 4 versus other scores for the transfer and mobility score (which ranges from 0 to 6) was data derived.

Their chosen model included the remaining five predictor variables. The authors gave these equal weight despite their clear unequal importance. The odds ratios ranged from 2.1 to 20.9, although the more relevant log-odds ratios varied by the smaller factor of 4. While there is a good case for simplicity in this study, there must be some consequent loss of predictive ability. As the magnitude of this effect is not reported, the case for not weighting is not fully justified. Further, no

Table V. Performance of the STRATIFY score on all three data sets using the authors' chosen cut-off (risk score ≥ 2) [40]. Values are proportions (and number of falls/number of patient risk assessments) in each subgroup.

Score	Original (case control) study ($n = 232$)		Local validation ($n = 395$)		External validation ($n = 442$)	
≥ 2	0.82	(78/95)	0.62	(66/106)	0.39	(73/188)
< 2	0.28	(38/137)	0.02	(5/289)	0.02	(6/254)
Total	0.50	(116/232)	0.18	(71/395)	0.18	(79/442)
PSEP (95 per cent CI)	0.54 (0.44 to 0.65)		0.61 (0.51 to 0.70)		0.36 (0.29 to 0.44)	

allowance was made for the inter-relationships between the variables, as the selection was based on univariate analyses. Finally, whereas the first study looked at all falls the two subsequent evaluations considered falls within the following week. It is not possible to judge the effect of these aspects of the study.

Table V shows the performance of the STRATIFY score using the authors' chosen cut-off on all three data sets. The properties of the score were notably different in the two prospective evaluations, partly reflecting a lower prevalence of falls. The positive predictive value (p_{worst}) of a high score in the external evaluation (0.39) was less than half the value in the original study (0.82). PSEP actually increased in the local validation but was lowest in the external validation.

Despite concerns that the model is not the best that can be found, and that it is not statistically valid, the external validation of the STRATIFY risk score suggests that it has clinical value. Nevertheless, further evaluation elsewhere is certainly desirable and we might expect that other models could do even better.

6. DISCUSSION

6.1. Lessons from the case studies

Although no attempt was made to ensure that the case studies were representative of the wider medical literature, certain lessons may be drawn from them.

It seems that authors tend to confirm the validity of their own models, but that others are less successful in doing so. Our explanations are speculative. Publication bias may be one factor. If authors fail to validate their model successfully, it is doubtful that they would attempt to publish the results. Second, the test set of patients is often drawn from the same centre as the original set, but at a later time. Clearly there will be many similarities between the two sets of patients and between the clinical and laboratory techniques used in evaluating them. This will pose a lesser challenge to a prognostic model than a new setting with different investigators and techniques, and perhaps a more sceptical approach. A dramatic example is the prediction of relapse in asthma (Section 5.2) in which the original model essentially failed completely when evaluated independently. The falls example (see Section 5.4) shows a similar tendency in that the performance of the prognostic index held up quite well in the validation study in the original hospital but was much

less effective in the remote hospital. Third, if the original model and its evaluation with new data are published by the same investigators in a single article, there may be a temptation, if the model does not fit the new data too well, to refine it in the light of the new data and publish it as though it was the original model. We have no evidence that such an approach has been applied, but it does not seem totally improbable. Our general conclusion here is congruent with normal practice in 'hard science': to be acceptable, a theory (model) must be confirmed by different workers in a different setting. As Laupacis *et al.* note, 'It is essential to prospectively validate the rule in a group of patients different from the group in which it was derived, preferably with different clinicians' [41].

Many authors seem unclear about (or at least do not state clearly) the aims of the prognostic model they are developing. It is as though 'demonstrating prognostic value' is deemed sufficient in itself; worse, it is widely believed that such a demonstration may be based on the statistical significance of predictors in a multivariable model. Similarly, and perhaps consequently, when evaluating a model with new data authors seem instinctively 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. The evaluation of the coronary prognostic index by Woo *et al.* [39] (see Section 5.3) is an example. This is a pity since their index happens to perform well by more relevant criteria such as the ability to reproduce the value of PSEP adequately in new patients. Of course when authors' aims are more specific, as were those of Gibson and Stephenson [33] (see Section 5.1), they make themselves more vulnerable to failure. In general authors show no appreciation of a distinction between statistically and clinically valid models.

It is striking that the statistical problem of overoptimistic prediction is mentioned in very few prognostic studies and certainly in none of our case studies. There may be several reasons. First, although the idea is now much more widely known to statisticians, the technical developments are fairly recent and may have not permeated the statistical community fully, let alone the medical research community. Second, although statistical techniques such as shrinkage and bootstrapping are available to attempt to reduce overoptimism at the model building stage, they tend to be computer intensive. As yet, there is no evidence as to their effectiveness in accurately estimating the prediction error seen in an 'ideal' validation setting with data collected by different investigators in a different centre. Third, the application of the techniques may make the research findings less impressive, even, in extreme cases, reducing what appeared to be a useful predictive model to worthlessness. We believe that over-optimism was the main contributor to the failure of the model of Fischl *et al.* [35] in the hands of Centor *et al.* [37] (see Section 5.2). The model was selected by stepwise discriminant analysis with 14 candidate variables, followed by data-driven 'optimal' cutpoint selection. However, there were only 40 events (relapses) among the 160 patients used to develop the model.

6.2. General points

An important question which arises in an attempt to validate a model with new data is 'what is the final model?'. We have seen and should expect that in most cases perfect 'validity' will not be achieved. van Houwelingen and Thorogood [42] took the view that the original model should be updated in the light of the new data, and proposed a way of doing that in a survival analysis setting. Their updating method involved a minor calibration of the prognostic index from the original model, rather than a complete reconstruction. From the values of the calibration parameters they inferred that prognosis for the event in question (graft failure or death following kidney transplantation) had somewhat improved over time. We think such calibration may be

sensible where the original model has been carefully built and seems to predict fairly well. However, many other types of 'failure to validate' are conceivable, and in some cases the original model may be too flawed to salvage. If the case-mix in the validation sample differs in a major way from that of the construction sample, the model may fail. However, it may be possible to improve it by including new variable(s) which relate to the different case-mix and are found to be prognostic in the new sample. For example, the range of patients' ages in the construction and validation samples might differ markedly, resulting in age not being recognized initially as an important prognostic factor.

Another important issue is that of simplicity versus complexity of a final model, and its effect on transportability. Occam's razor suggests a preference for smaller rather than larger models if a 'true' small model exists, but whether there really is a 'true' model is a philosophical question. Experience indicates that a larger model is more likely to give overoptimistic predictions when extensive variable selection has been performed [10]. This was the case in van Houwelingen and Thorogood's study [42] in which a separate parameter was fitted for each of 52 transplantation centres supplying data. However, rather than remove centres from the model, they used empirical Bayes methods to obtain shrunken estimates of the regression coefficients for the centres, which resulted in improved predictive performance. Although shrinkage-based approaches to reduce overoptimism have appeal (and the application to kidney transplantation is a good example), practicalities must be considered. For instance, a specific requirement of the coronary prognostic index for the Chinese [38] was that it be 'simple, practical and economic', meaning that sophisticated and expensive measurement techniques not likely to be available in Chinese hospitals at that time were disallowed. Similar considerations applied to Oliver *et al.*'s prognostic index [40] for propensity to fall discussed in Section 5.4. In general and quite apart from statistical arguments, simplicity of models and reliability of measurements are clearly important criteria in developing clinically useful prognostic models [22]. The more philosophical issues of simplicity versus complexity will no doubt continue to be debated and exemplified within the context of given data sets and analyses.

Rather like validation itself, the creation of prognostic groups is a process that is much done but seldom discussed. If a prognostic index with many possible values is obtained, it is not obvious how to decide the number and nature of the groups. We suggest that since the aim of prognostic studies is to create clinically valuable indexes, the definition of groups should be driven mainly by clinical rather than statistical criteria, particularly in the case of pragmatic studies. For example, suppose a clinician would leave untreated a patient with at least a 90 per cent chance of surviving five years, would apply aggressive therapy if the prognosis was 30 per cent survival or less, and would use standard therapy in intermediate cases. Three prognostic groups would appear to be sensible, and validation of the model would address whether such large separation ($PSEP = 0.6$) was reliably available to other investigators in other settings. In explanatory studies the question of how best to define prognostic groups (or even whether to produce groups) remains an open question. Of course some techniques such as CART or very simple prognostic models involving few combinations of factor values generate groups directly. However, CART is not designed to generate clinically useful prognostic groups.

6.3. Quantifying performance

Two aspects of performance may be distinguished: (i) the amount of prognostic information, which relates to potential clinical usefulness; and (ii) whether the model performs as intended with

new data. There are many possible ways to define the amount of prognostic information, some of them discussed by Graf *et al.* [23]. A general measure applicable to any model is the Kullback–Leibler information [43], which tends to infinity as the predictive ability increases. We reiterate that our probability measure, PSEP, is not meant to be a definitive proposal, but as we hope we have shown, it has the merit of simplicity and interpretability. Hand [44] considered several properties of classification estimators and ways to evaluate their performance in a predictive setting. It would be desirable to investigate the properties of PSEP with respect to some of Hand's measures.

With the data set on which the model was developed, PSEP is a measure not of the amount of prognostic information in the data but of the *estimated* ability of a rule to separate individuals into prognostic groups. Analysis of an independent data set will show if the estimate was reasonable. We found it useful when describing the results of the analysis of prognostic data in the case studies. When there are more than two groups it is important to ensure that the extreme groups, from which PSEP is calculated, are large enough and have sufficient events to allow adequate estimates of p_{worst} and p_{best} . If not, amalgamation of groups may be required. In survival studies PSEP is a function of follow-up time. Typically the clinician will have in mind a survival time (or times) of interest, and PSEP will be estimated from the difference between the Kaplan–Meier survival curves for the extreme prognostic groups at the each time.

PSEP is of course the difference between two proportions, also known as the risk difference in the context of randomized trials. The reciprocal of the risk difference is the 'number needed to treat', a measure increasingly being used in reporting trial results [45]. It represents the number of a patients that need to receive a new treatment to prevent an additional adverse outcome. By analogy the reciprocal of PSEP represents the number of extra patients in the high risk group needed to generate one additional adverse event. So, for example, in the STRATIFY example above (Section 5.4), PSEP was 0.36 in the external validation sample and its reciprocal is about 3. So, using this estimate we can say that if three patients were assessed there would be one additional fall if they were all high risk patients than if they were all low risk patients.

6.4. Concluding comments

A vast number of prognostic models are published each year, yet, although they have considerable potential clinical value, prognostic indices are not widely used in clinical practice. We believe this is most probably because most have not been demonstrated to be effective by other investigators in other centres [22]. To be useful, a prognostic index should be clinically credible, accurate, have generality (that is, be validated elsewhere), and, ideally, be demonstrated to be clinically effective, in the sense that it provides useful additional information to clinicians [22].

In addition, the study should be described in adequate detail. Concato *et al.* [3] reviewed 44 publications which presented prognostic models. They found frequent methodological shortcomings, including unspecified method for selecting variables in the model, unspecified coding of variables, and risk of overfitting through too few events per variable. Other authors have also found deficiencies in both reporting and methodology used [9, 41].

In a useful and carefully argued recent paper, Justice *et al.* [46] explored the assessment of the effectiveness of a prognostic system in some detail. They discussed measures of accuracy (calibration and discrimination) and distinguished between reproducibility and transportability. Reproducibility is defined as the performance of a model on a sample of similar patients not included in the development of the model. Transportability reflects a model's ability to predict

among patients from different but 'plausibly related' populations. The authors proposed (see their Table 2) a systematic and increasingly stringent set of criteria for model transportability comprising resistance to variations in historical period (essentially, what we term temporal validation), geographic location, measurement methodology, disease severity, and follow-up interval. They used as examples the development of the Dukes and Jass staging systems in colorectal cancer. Whether their hierarchy of validation tests will prove to be useful in other clinical settings remains to be seen.

Given that the aim of developing a prognostic model (in what we have referred to as a pragmatic study) is the prediction of outcome for future patients, it seems clear that this objective should be kept in mind when producing a model. The evaluation of the performance of a model will be much simplified when it is only a check that the model does in fact work – rather like comparing the characteristics of treatment groups in a randomized trial. It follows that study design and model building should anticipate and deal with overoptimism. Further studies are needed to ascertain which design features and analysis procedures are likely to lead to a good model.

ACKNOWLEDGEMENTS

We thank Hans van Houwelingen, Willi Sauerbrei and Martin Schumacher (the other members of the COSMIC group), Marc Buyse and Jeremy Wyatt for helpful comments and discussions.

REFERENCES

1. Feinstein AR. *Multivariable Analysis: an Introduction*. Yale University Press: New Haven, 1996; 184–187, 578–582.
2. Diamond GA. Future imperfect: the limitations of clinical prediction models and the limits of clinical prediction. *Journal of the American College of Cardiology* 1992; **14**:12A–22A.
3. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Annals of Internal Medicine* 1993; **118**:201–210.
4. Chatfield C. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A* 1995; **158**:419–466.
5. Burstein AH. Fracture classification systems: do they work and are they useful? *Journal of Bone and Joint Surgery* 1993; **75A**:1743–1744.
6. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and accuracy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
7. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. *Journal of Clinical Epidemiology* 1997; **50**:21–29.
8. Streiner DL, Norman GR. *Health Measurement Scales: a Practical Guide to their Development*. 2nd edn, Oxford University Press: New York, 1995.
9. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Statistics in Medicine* 1995; **14**:331–345.
10. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics*, 1999; **48**:313–329.
11. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic modelling. *Statistics in Medicine* 1984; **3**:143–152.
12. Gamel JW, McCurdy JB, McLean IW. A comparison of prognostic covariates for uveal melanoma. *Investigative Ophthalmology and Visual Science* 1992; **33**:1919–1922.
13. Roecker EB. Prediction error and its estimation for subset-selected models. *Technometrics* 1991; **33**:459–468.
14. Penny W, Frost D. Neural networks in clinical medicine. *Medical Decision Making*, 1996; **16**:386–398.
15. Ohno-Machado L. A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. *Computers in Biology and Medicine* 1997; **27**:55–65.
16. Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research* 1993; **26**:74–97.

17. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. *Statistics in Medicine* 1998; **17**:2501–2508.
18. Gray-Donald K, Kramer MS. Causality inference in observational vs. experimental studies. An empirical comparison. *American Journal of Epidemiology* 1988; **127**:885–892.
19. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 1994; **6**:979–985.
20. Peduzzi P, Concato J, Feinstein AR, Holford TR. The importance of events per independent variable (EPV) in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology* 1995; **48**:1503–1510.
21. Peduzzi P, Concato J, Kemper E, Holford TR, and Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
22. Wyatt, JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal* 1995; **311**:1539–1541.
23. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**:2529–2545.
24. Cox DR. A note on data-splitting for the evaluation of significance levels. *Biometrika* 1975; **62**:441–444.
25. Hirsch RP. Validation samples. *Biometrics* 1991; **47**:1193–1194.
26. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman and Hall: London, 1993; 255.
27. Verweij PJM, van Houwelingen JC. Cross-validation in survival analysis. *Statistics in Medicine*. 1993; **12**:2305–2315.
28. Schumacher M, Holländer N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Statistics in Medicine* 1997; **16**:2813–2827.
29. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Statistics in Medicine* 1991; **10**:1213–1226.
30. Brier GW. Verification of weather forecasts expressed in terms of probability. *Monthly Weather Review* 1950; **78**:1–3.
31. Efron B. Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association* 1978; **73**:113–121.
32. Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Annals of Internal Medicine* 1996; **125**:406–412.
33. Gibson RM, Stephenson GC. Aggressive management of severe closed head injury: time for reappraisal. *Lancet* 1989; **ii**:369–371.
34. Feldman Z, Contant CF, Robertson CS, Narayan RK, Grossman RG. Evaluation of the Leeds prognostic score for severe head injury. *Lancet* 1991; **337**:1451–1453.
35. Fischl MA, Pitchenik A, Gardner LB. An index predicting relapse and need for hospitalization in patients with acute bronchial asthma. *New England Journal of Medicine* 1981; **305**:783–789.
36. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**:829–835.
37. Centor RM, Yarbrough B, Wood JP. Inability to predict relapse in acute asthma. *New England Journal of Medicine* 1984; **310**:577–580.
38. Woo K-S. Coronary prognostic index for the Chinese. *Australia and New Zealand Journal of Medicine* 1987; **17**:562–567.
39. Woo K-S, Pun CO, Want RYC, Ma H, Huang ZZ, Dai RH, Huang DJ, Vallance-Owen J. Validation of a coronary index for the Chinese—a tale of three cities. *International Journal of Cardiology*, 1989; **23**:173–178.
40. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. *British Medical Journal* 1997; **315**:1049–1053.
41. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *Journal of the American Medical Association* 1997; **277**:488–494.
42. van Houwelingen JC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**:1999–2008.
43. Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79–86.
44. Hand DJ. *Classification and Assessment of Classification Rules*. Wiley: Chichester, 1997.
45. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal* 1995; **310**:452–454.
46. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**:515–524.