STATISTICS IN MEDICINE, VOL. 10, 1213–1226 (1991)

VALIDATION TECHNIQUES FOR LOGISTIC REGRESSION MODELS

MICHAEL E. MILLER AND SIU L. HUI

Division of Biostatistics, Indiana University Department of Medicine, and the Regenstrief Institute for Health Care, Riley Research Wing, Rm 135, 702 Barnhill Drive, Indianapolis, IN 46202-5200, U.S.A.

AND

WILLIAM M. TIERNEY

Division of General Internal Medicine, Indiana University Department of Medicine, and the Regenstrief Institute for Health Care, 1001 W. 10th St., Indianapolis, IN 46202, U.S.A.

SUMMARY

This paper presents a comprehensive approach to the validation of logistic prediction models. It reviews measures of overall goodness-of-fit, and indices of calibration and refinement. Using a model-based approach developed by Cox, we adapt logistic regression diagnostic techniques for use in model validation. This allows identification of problematic predictor variables in the prediction model as well as influential observations in the validation data that adversely affect the fit of the model. In appropriate situations, recommendations are made for correction of models that provide poor fit.

1. INTRODUCTION

Statistical models developed for prediction need validation. In cross-validation, one divides the data randomly into two portions, one for model development and the other for model validation. Validation techniques are also required when investigators wish to ascertain whether a model developed at some other site (for example, hospital or medical centre) is appropriate for prediction in their particular setting; we call this external validation. Lastly, in temporal validation, one validates prospectively a model developed at a specific point in time within the same setting at some future date. External and temporal validation are the primary focus of this paper.

When validating logistic regression models, the major question typically concerns how well the predicted probabilities agree with the responses within the independent sample. A goodness-of-fit statistic provides a summary measure of the deviations of individual predicted probabilities from the actual outcomes. The total deviation can come from three major, albeit non-exhaustive, sources. The first arises if on average the predicted probability of a positive result in the validation set is too high or too low. This may happen with application of the validation at a new site that has a higher or lower prevalence of outcomes than that at the site of model development. This component of predictive accuracy has been referred to as 'calibration' or 'reliability'.[1,2]

A second source of deviation reflects a model's 'refinement' or 'spread' and addresses how well the model can discriminate between observations that have positive and negative outcomes. In model validation, one primary concern is whether the discriminatory ability obtained in the

0277–6715/91/081213–14$07.00

Received August 1990
Revised January 1991

developmental data set diminishes in the validation set. This often occurs simply because the first estimate of discrimination comes from the same data that were used to fit the model.[3] The decline in discrimination, however, may arise from the inclusion in the model of some non-predictive variables. In this situation, poor refinement may result either from type I error in model fitting or from application of the model in a different setting.

While calibration and refinement are systematic sources of predictive inaccuracy, a third source may be subsets of observations in the validation set for which the model does not perform well; that is, their contribution to the overall deviation from the model is exceptionally high. These observations are analogous to the influential observations in the diagnostics of model building. There is little in the literature on how to identify these observations or what to do when the fit is unsatisfactory.

Since an unbiased estimate of the prediction error is the main goal of cross-validation, modifying the model, for example by adding or deleting covariates, invalidates the assessment of fit.[4] However, there are situations where an externally or previously derived model is to be adapted for contemporary use locally. If the model is found not to fit well, the local organization may not have the authority, resources or adequate sample size for the development of a new model. For such situations, we propose some techniques to identify the problems causing poor fit, and to correct them with minimal adjustments to the model. For example, a change in the outcome's prevalence requires only an adjustment to the intercept. Although not described herein, our validation approach can also be extended for use with non-binary outcomes (such as linear models with continuous dependent variables).

In this paper, we present a comprehensive approach to model validation for binary outcomes using an independent sample. Section 2 reviews the various measures of goodness-of-fit and its three components. In particular, we draw attention to a little known approach developed by Cox[1] that provides summary measures of a model's calibration and refinement, and has attracted renewed interest in recent years.[5–7] In Section 3, we extend Cox's approach to allow application of diagnostic procedures to model validation. We then propose a strategy for model validation that can detect both 'influential' observations and problem predictors in the model, and that subsequently allows for adjustments to the model. We demonstrate this strategy with an example that illustrates validation of a prediction model for hypokalaemia (low blood potassium concentration).

2. REVIEW OF THE LITERATURE

2.1. Measures for overall goodness-of-fit

Most goodness-of-fit statistics apply to model fitting; that is, they measure overall model fit with the same data used for model development. For logistic regression models, Efron constructs a wide class of measures that includes counting error (proportion misclassified), squared error, and deviance.[3] All of these measures are based on some measure of deviation of the fitted model from the saturated model and can be used in model validation. Van Houwelingen and Le Cessie give details for a number of specific examples.[4]

When a validation sample is available, let i index the N unique profiles, each containing n_i observations, defined by the cross-classification of the levels of all covariates (continuous or discrete). Application of the fitted model to the validation sample results in the predicted probabilities $\pi_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}_d)/[1 + \exp(\mathbf{x}_i'\boldsymbol{\beta}_d)]$, where $\boldsymbol{\beta}_d$ is a $p \times 1$ vector of assumed population parameters estimated from the developmental sample, and $\mathbf{x}_i'$ is the ith row of the $N \times p$ matrix $\mathbf{X}$ that contains the covariates from the validation sample. A common measure of squared error is

the chi-square goodness-of-fit statistic. When $n_i\pi_i$ is sufficiently large to use the normal approximation to the binomial distribution, we can assess the overall fit of the model with a Pearson chi-square statistic

$$\chi^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{n_i\pi_i(1 - \pi_i)}, \tag{1}$$

where o_i denotes the number of positive responses out of the n_i observations in the ith profile, and $e_i = n_i\pi_i$ represents the expected number of responses. With the parameter estimates from the predictive model assumed known, we say the model fits the validation sample poorly if χ^2 greatly exceeds its degrees of freedom N. When the expected numbers in the cells are small, Lemeshow and Hosmer[8] and Hosmer and Lemeshow[9] suggest computation of (1) based on an appropriate grouping of the observations, that is replacing the π_i in (1) by π_k, the average probability within the kth group, $k = 1, \ldots, g$. One preferred approach is to set $g = 10$ and compute (1) using deciles of risk.

The deviance provides another overall measure of the fit of a predictive logistic model to a validation sample. Its interpretation is as the likelihood ratio test statistic of a saturated model with N estimated parameters versus a known model with zero estimated parameters. Thus, as applied to model validation, if the deviance substantially exceeds N then we say the model fits poorly.

2.2. Measures of calibration

A model is well calibrated if the average of the predictive probabilities equals the proportion of positive outcomes in the sample. It is always attained with validation performed on the model development data. Harrell *et al.* argue that calibration is of secondary importance because, in the presence of good discrimination, bad calibration is correctable.[2] In this case, however, the overall goodness-of-fit measures may show a very poor fit for the model even when there is a small but easily correctable calibration bias. For example, if we calculate (1) based on deciles of risk, where $n_i = n$ and the observed proportions are consistently higher than the expected by a factor k (that is $o_i = ke_i$), then $\chi^2 = n(k-1)^2\sum[\pi_i/(1-\pi_i)]$ can be large when k is close to 1 but n is large. Cox provides an approach for the detection of calibration bias.[1]

2.3. Measures of refinement

The measurement of refinement is conceptually a more difficult task than the measurement of calibration because the ability of a predictive model to spread probability estimates depends on the observed population of elements. Quite conceivably, the true distribution of underlying probabilities in the validation set may cover only a restricted range of the unit interval. Consequently, for a measure of refinement to be useful, not only should it describe the discriminatory ability of the predictor in relation to the outcomes, but also it should not be greatly affected by the underlying range of probabilities.

Any parametric measure of refinement, such as the point-biserial correlation,[10] would have its distribution dependent on the underlying distribution of the predictive probabilities. To avoid this problem, Harrell *et al.* proposed a non-parametric measurement of refinement that uses only the rank of the magnitude of the assessment error.[11] Harrell *et al.*'s c-index measures the probability that for two randomly chosen patients, one with and the other without disease, the one with the higher probabilistic prediction has the outcome of interest.[11] This index relates directly to the area under a 'receiver operating characteristic curve' (ROC), and can be obtained as the parameter of the Mann–Whitney–Wilcoxon rank sum test.[12] Thus the c-index does not

depend on interpretation of the assessments as probabilities, and we can obtain the same c-index by applying a monotonic transformation to the assessed probabilities. Proposals for several extensions of ROC methodology have appeared for comparison of two or more areas of ROC curves.[13–16] One can use some of these comparison methods to test whether the refinement measure obtained in the developmental data set alters substantially when one applies the model to a validation set.

2.4. Cox's measures of calibration and refinement

Cox proposed an approach for the assessment of probability assessors that permits a greater understanding of the differing degrees of refinement.[1] This approach, directly applicable to model validation, uses logistic regression for testing the agreement between a series of hypothesized probabilities and a binary outcome variable. The approach assumes that $Y_1, Y_2, \ldots, Y_n$ are mutually independent random variables that take the values 0 or 1, each Y_i having an associated π_i specified by the model, $0 \leqslant \pi_i \leqslant 1$. The goal of the analysis is to test the hypothesis

$$p_i = Pr(Y_i = 1) = \pi_i, \qquad i = 1, \ldots, n. \tag{2}$$

Cox places a logistic model on the logit of $Pr(Y_i = 1)$ to allow a test of (2). This model takes the form

$$\log[Pr(Y_i = 1)/Pr(Y_i = 0)] = \alpha + \beta \log[\pi_i/(1 - \pi_i)], \tag{3}$$

where a test of H_0: $\alpha = 0$ and $\beta = 1$ is a test of the hypothesis stated in (2). If $\beta > 1$, the π_i show the correct direction but do not vary enough; if $0 < \beta < 1$, the π_i vary too much; if $\beta < 0$, the π_i show the wrong general direction; and if $\beta = -1$, the π_i are the complements of the true probabilities p_i.

At $\pi_i = 0{\cdot}5$, $\log[Pr(Y_i = 1)/Pr(Y_i = 0)] = \alpha$, and $\alpha = 0$ implies $Pr(Y_i = 1) = 0{\cdot}5$. Thus α is a parameter that reflects the overall calibration of the model if $\beta = 1$, and the calibration at $\pi_i = 0{\cdot}5$ if $\beta \neq 1$. The predictive probability is generally too low if $\alpha > 0$ and too high if $\alpha < 0$. Cox presents several score tests appropriate for testing hypotheses related to model (3). Three important hypotheses testable with use of either score or likelihood ratio tests are: (1) H_0: $\alpha = 0$, $\beta = 1$, an overall test for the reliability of the predictions; (2) H_0: $\alpha = 0 \mid \beta = 1$, a test of an incorrect calibration given appropriate refinement; and (3) H_0: $\beta = 1 \mid \alpha$, a test of refinement given correct calibration. The third hypothesis is due to Harrell and Lee,[7] who provide a lucid discussion of the above hypotheses and present several indices of refinement that result from various decompositions of the likelihood. These authors also investigated the power associated with Cox's tests of reliability.[7]

Recently, Bloch[6] related Cox's approach to various indices for assessing probability assessors, e.g. the Brier score[17] and Shapiro's Q index.[18] Bloch illustrated that the Brier score is highly correlated with Shapiro's Q, and several authors have shown that Cox's statistic for testing H_0: $\beta = 1 \mid \alpha = 0$ is equivalent to Shapiro's Q.[5–7] Since Shapiro's index measures refinement with the assumption of perfect calibration, Bloch suggests that neither Shapiro's Q nor the Brier score is an appropriate index in the presence of poor calibration.

3. A STRATEGY FOR EXTERNAL AND TEMPORAL MODEL VALIDATION

In any model validation exercise, one should screen the validation sample to identify observations that contain covariate values outside the range observed in the developmental sample. Subsequently, one may choose to exclude these observations from the validation sample to avoid

extrapolation of the model to subpopulations of subjects for whom it was not explicitly developed. However, if desired, the methods we suggest do allow one to ascertain whether the model is appropriate within these subpopulations.

To obtain an overall measure of the goodness-of-fit of the model, we can calculate one of the appropriate statistics described in Section 2.1. One should, however, pay attention not only to the overall statistic, but also to inspection of the individual deviations that provide an indication of which covariate profiles the model does not fit well. Furthermore, one can identify poor calibration and/or refinement by inspection of a plot of the observed versus the expected proportions. Deviation from the 45° line of identity (representing $\alpha = 0$, $\beta = 1$) indicates poor calibration and/or refinement, as discussed in Section 2.4.

One can use either score or likelihood ratio tests to investigate hypotheses 1–3 specified in Section 2.4. For a further understanding of the model's fit to the validation sample, we suggest an extended application of Cox's model (3) to identify the influence of extreme observations on calibration and refinement. Once again, assume that the validation sample consists of N independent binomial random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$, where $Y_i \sim B(n_i, p_i)$. Furthermore, let $\mathbf{X}_v$ denote an $N \times 2$ matrix with ith row $\mathbf{x}_{iv}' = (1, L_i)$, where $L_i = \log[\pi_i/(1 - \pi_i)]$. Using this notation, we can write Cox's model (3) as $\boldsymbol{\theta} = \mathbf{X}_v\boldsymbol{\beta}_v$, where $\boldsymbol{\beta}_v = [\alpha\ \beta]'$, and $\boldsymbol{\theta}$ is the vector that contains the terms $\theta_i = \log[p_i/(1 - p_i)]$. Following Pregibon[19] and McCullagh and Nelder,[20] we can obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_v$ with use of iteratively reweighted least squares, the $(k + 1)$th estimate obtained by iterating the equation

$$\hat{\boldsymbol{\beta}}_v^{(k+1)} = (\mathbf{X}_v'\mathbf{V}^{(k)}\mathbf{X}_v)^{-1}\mathbf{X}_v'\mathbf{V}^{(k)}\mathbf{z}^{(k)},$$

where $\mathbf{z}^{(k)} = \mathbf{X}_v\hat{\boldsymbol{\beta}}_v^{(k)} + [\mathbf{V}^{(k)}]^{-1}\mathbf{s}^{(k)}$, $\mathbf{V}^{(k)}$ is an $N \times N$ diagonal matrix with diagonal elements $v_{ii} = n_i\hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})$, $\mathbf{s}^{(k)} = (s_1^{(k)}, \ldots, s_N^{(k)})'$ and $s_i^{(k)} = y_i - n_i\hat{p}_i^{(k)}$.

This model-based approach to validation of logistic regression models allows the application of regression diagnostics[19,21] for identification of data points within the validation sample that exert undue influence on the estimated calibration (α) and refinement (β) terms. Define $\hat{\boldsymbol{\beta}}_{iv} = [\hat{\alpha}_i, \hat{\beta}_i]'$ as the vector that contains estimates of calibration and refinement with exclusion of the ith profile from the fitting process. The general formula for a one-step estimator of $\hat{\boldsymbol{\beta}}_{iv}$ appears in Cook and Weisberg[21] as

$$\hat{\boldsymbol{\beta}}_{iv} = \hat{\boldsymbol{\beta}}_v - \frac{(\mathbf{X}_v\mathbf{V}\mathbf{X}_v)^{-1}\mathbf{x}_{iv}s_i}{1 - h_{ii}},$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$. At convergence, the matrix $\mathbf{H}$ takes the form

$$\mathbf{H} = \mathbf{V}^{(1/2)}\mathbf{X}_v(\mathbf{X}_v'\mathbf{V}\mathbf{X}_v)^{-1}\mathbf{X}_v'\mathbf{V}^{(1/2)},$$

and in this application we can express the ith diagonal element as

$$h_{ii} = \frac{v_{ii}(\sum_{j=1}^N v_{jj}L_j^2 - 2L_i\sum_{j=1}^N v_{jj}L_j + L_i^2\sum_{j=1}^N v_{jj})}{(\sum_{j=1}^N v_{jj})(\sum_{j=1}^N v_{jj}L_j^2) - (\sum_{j=1}^N v_{jj}L_j^2)^2}.$$

Since perfect calibration corresponds to $\alpha = 0$, one useful diagnostic measure would quantify the influence that the ith profile has in pulling the estimate of α away from zero. We express a measure of this influence as

$$\Delta\text{stat}\,\hat{\alpha}_i = \frac{(\hat{\alpha} - 0)^2}{\text{var}(\hat{\alpha})} - \frac{(\hat{\alpha}_i - 0)^2}{\text{var}(\hat{\alpha}_i)}, \tag{4}$$

where

$$\text{var}(\hat{\alpha}_i) = \frac{\sum_{j \neq i} v_{jj} L_j^2}{(\sum_{j \neq i} v_{jj})(\sum_{j \neq i} v_{jj} L_j^2) - (\sum_{j \neq i} v_{jj} L_j^2)^2},$$

and $v_{jj} = n_j \hat{p}_j(1 - \hat{p}_j)$ is obtained using $\hat{\alpha}_i$ and $\hat{\beta}_i$. Large positive values of $\Delta\text{stat}\,\hat{\alpha}_i$ indicate profiles associated with poor calibration; large negative values of $\Delta\text{stat}\,\hat{\alpha}_i$ correspond to profiles that, when dropped, make calibration become worse; and small values of $\Delta\text{stat}\,\hat{\alpha}_i$ indicate very little change in calibration as a result of dropping the profile. Likewise, we obtain a useful measure of a profile's influence on refinement by using

$$\Delta\text{stat}\,\hat{\beta}_i = \frac{(\hat{\beta} - 1)^2}{\text{var}(\hat{\beta})} - \frac{(\hat{\beta}_i - 1)^2}{\text{var}(\hat{\beta}_i)}, \tag{5}$$

where

$$\text{var}(\hat{\beta}_i) = \frac{\sum_{j \neq i} v_{jj}}{(\sum_{j \neq i} v_{jj})(\sum_{j \neq i} v_{jj} L_j^2) - (\sum_{j \neq i} v_{jj} L_j^2)^2}.$$

Interpretation of values of $\Delta\text{stat}\,\hat{\beta}_i$ is similar to that described for $\Delta\text{stat}\,\hat{\alpha}_i$. Calculating $\text{var}(\hat{\alpha}_i)$ and $\text{var}(\hat{\beta}_i)$ can be burdensome when the number of profiles is large; thus we suggest that another informative statistic can be obtained by substituting $\text{var}(\hat{\alpha})$ and $\text{var}(\hat{\beta})$ for $\text{var}(\hat{\alpha}_i)$ and $\text{var}(\hat{\beta}_i)$ in equations (4) and (5), respectively. Making this substitution slightly changes the interpretation of (4) and (5) since the new statistic only measures changes associated with the estimated coefficients.

Plots of $\Delta\text{stat}\,\hat{\beta}_i$ and $\Delta\text{stat}\,\hat{\alpha}_i$ against the profile number help to identify profiles that contribute greatly to poor calibration and/or refinement. Additionally, univariate plots of $\Delta\text{stat}\,\hat{\beta}_i$ and $\Delta\text{stat}\,\hat{\alpha}_i$ against each covariate used in the predictive model help to identify extreme ranges of the continuous covariates that are poorly fitted by the model. With discrete covariates contained in the predictive model, one can perform the above plots with use of different symbols for each level of the covariate to assess the influence of these discrete factors.

The magnitude of $\Delta\text{stat}\,\hat{\beta}_i$ and/or $\Delta\text{stat}\,\hat{\alpha}_i$ will be partially influenced by the number of profiles in the validation data set. Thus we do not advocate strict rules regarding the classification of 'outliers'. Visually inspecting the plots can frequently provide an indication of those profiles that may alter the conclusions obtained from hypothesis tests directed at calibration and/or refinement. For instance, if $\hat{\alpha}^2/\text{var}(\hat{\alpha}) = 4{\cdot}0$, then profiles with $\Delta\text{stat}\,\hat{\alpha}_i > 0{\cdot}16$ $(4{\cdot}0 - 3{\cdot}84)$ might be of interest since deletion of these observations affects the overall conclusions regarding calibration. In contrast, one may encounter a situation where no individual profile changes the overall conclusions, but a group of profiles sharing a common characteristic all have small positive values for $\Delta\text{stat}\,\hat{\beta}_i$ and/or $\Delta\text{stat}\,\hat{\alpha}_i$. When 'outlying' profiles over-represent specific groups of subjects, then one can add additional terms to model (3) to investigate calibration and/or refinement within these subgroups. For instance, if age is a predictor variable and observations greater than the 90th percentile of the age variable generally seem to influence either calibration or refinement, we can then include in the model a dummy variable x_{dum} that identifies subjects who fall in the upper decile of age,

$$\theta_i = \alpha_1 + \beta_1 L_i + \alpha_2 x_{\text{dum}} + \beta_2 (x_{\text{dum}} * L_i) \tag{6}$$

to provide estimates and tests of the calibration and refinement within this group of subjects. If we cannot reject H_0: $\beta_2 = 0$, then the refinement within the subgroup does not appear to differ from that for the whole sample. Likewise, if we cannot reject H_0: $\alpha_2 = 0$, then the data are consistent with the hypothesis of equal calibration within both groups.

When the validation procedures have identified subsets of subjects for whom the model does not fit well, and the model is to be used for future prediction, then we must either adjust the model or place restrictions on its use. In contrast to cross-validation,[4] if we reject H_0: $\beta = 1 \mid \alpha$ we generally recommend considering the information gained from the proposed diagnostic methods before attempts to correct the model by multiplying all of the regression coefficients in the original logistic prediction model by $\hat{\beta}$. After identification of a poorly fitted subgroup, we suggest use of only simple corrections, thus leaving the original model as intact as possible. The form of these corrections may be as simple as dropping a variable from the model or adjusting a single parameter within a subpopulation. If, after the correction, we cannot reject the three reliability hypotheses specified in Section 2.4, then we could use the corrected model for future observations.

4. APPLICATION TO A LARGE DATA SET

Tierney *et al.* used logistic regression to identify risk factors for hypokalaemia (serum potassium $< 3{\cdot}5$ mmol/L in a population of 5817 outpatients on chronic diuretic therapy.[22] These data were retrieved from the computerized Regenstrief Medical Record System,[23] and the model was derived on a randomly selected subset of 3833 patients (two-thirds of the original sample). Cross-validation was performed on the other one-third of the patients in the original sample, and temporal validation was undertaken on 2601 patients who received diuretics during the nine months following the original data retrieval. Using ROC curve methodology,[12] these authors indicated that the predictive equation exhibited good predictive power in both validation samples. The eight variables identified as predictors of hypokalaemia, the associated ranges for these variables, and the estimated regression coefficients appear in Table I. The variables listed in Table I, other than those that are self-explanatory, are defined as follows: 'mean K^+' is the mean of all potassium concentrations measured prior to the potassium value of interest; 'K^+ sparer use' identifies the concurrent use of a diuretic that helps protect against potassium loss in the urine; 'thiazide use' indicates the concurrent usage of hydrochlorothiazide, chlorothiazide, chlorthalidone or metolazone; and 'prior K^+' contains the most recent serum potassium concentration prior to the potassium value of interest.

To validate this predictive model (referred to as PM1) in the same outpatient clinic several years after its development, we retrieved data between June 1988 and June 1990 for 3737 outpatients on chronic diuretic therapy. Restriction of the observations to the originally observed variable ranges resulted in a final validation data set that contained 2364 patients. Application of the Hosmer–Lemeshow statistic to these data provided a chi-square of 50·72 on 10 degrees of freedom (d.f.), and indicated a poor fit of the model to the validation sample.

When we applied Cox's model (3) to these data, the estimates of α and β, respectively, were $-0{\cdot}58$ (SE = 0·100) and 0·86 (SE = 0·063). Table II contains results obtained from the likelihood ratio tests used to investigate hypotheses 1–3 specified in Section 2.4. These results indicate that PM1 is not appropriate for current use within our outpatient clinic. Figure 1 contains a plot of each patient's influence on calibration as defined by equation (4). Likewise, Figure 2 contains a similar plot for each patient's influence on refinement as calculated using equation (5).

Figures 1 and 2 indicate that the magnitude of influence on $\hat{\alpha}$ attributable to individual observations is somewhat greater than the influence on $\hat{\beta}$. We constructed plots of the influence measures against each predictor variable to investigate relationships between the influence measures and the predictor variables. Figure 3 shows one such plot for calibration. For this plot, we sorted the observations initially by thiazide use and then by mean K^+ within thiazide groups. Thus the first 1116 cases represent patients without a history of thiazide use, with case 1 having an observed mean K^+ less than or equal to that value observed for case 2, and so on. This plot

Table I. Predictive model for hypokalaemia[22]

Variable	Range	Parameter estimate
Mean K^+	2·77–5·68 mmol/L	− 1·7132
K^+ sparer use	0, 1	− 0·8533
Thiazide use	0, 1	0·7220
Patient age	18–103 years	− 0·0149
Time on K^+ sparer	0–9·4 years	0·1532
Prior K^+	2·3–6·5 mmol/L	− 0·4281
Last serum bicarbonate	10–44 mmol/L	0·0453
Last serum sodium	113–160 mmol/L	−,0·0385
Constant		11·3172

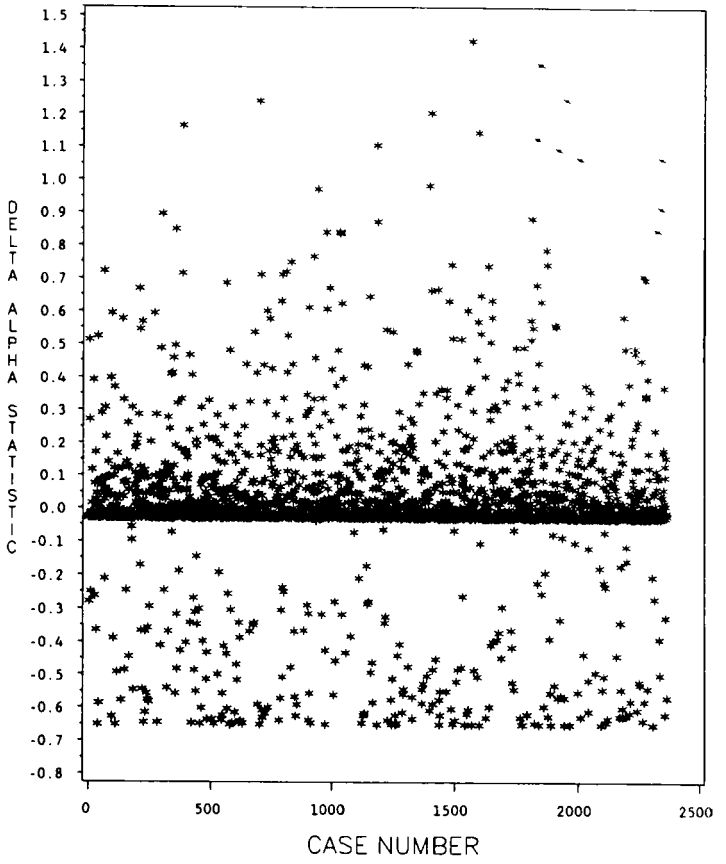

Figure 1. Initial diagnostic plot for alpha

displays several relationships: (a) patients with a low value of mean K^+ appear to have great influence on calibration, and (b) the predictive model appears to perform worse for patients with a history of thiazide use compared with non-thiazide users. Since $\hat{\alpha}^2/\text{var}(\hat{\alpha}) = 33{\cdot}46$, and the largest $\Delta\text{stat}\hat{\alpha}_i < 1{\cdot}5$, we can see that deletion of any individual profile does not alter our conclusions

Table II. Results of likelihood ratio tests for reliability

Null hypothesis	Test	Chi-square	d.f.	p-value
H_0: $\alpha = 0, \beta = 1$	$L(0, 1) - L(\alpha, \beta)$	47·38	2	$< 0{\cdot}01$
H_0: $\alpha = 0 \mid \beta = 1$	$L(0, 1) - L(\alpha, 1)$	42·61	1	$< 0{\cdot}01$
H_0: $\beta = 1 \mid \alpha$	$L(\alpha, 1) - L(\alpha, \beta)$	4·77	1	0·03

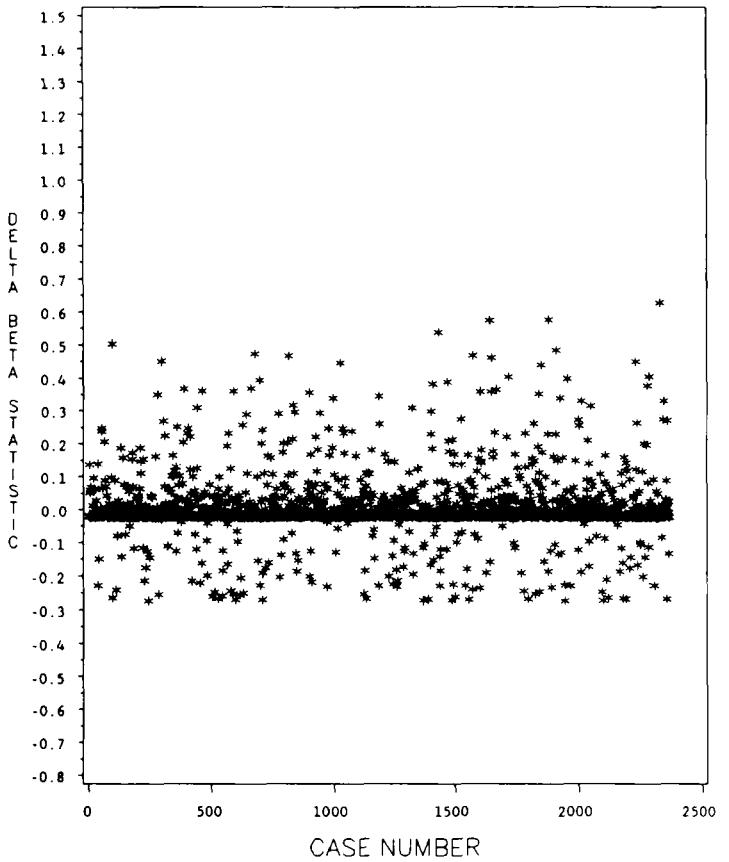

Figure 2. Initial diagnostic plot for beta

regarding calibration. However, if we arbitrarily define 'outliers' that contribute to poor calibration as observations with measures greater than 0·25 for $\Delta\text{stat}\hat{\alpha}_i$, we find that the 189 α 'outliers' greatly over-represent the group of patients with a history of thiazide use. In fact, within this group 78 per cent of patients have a history of thiazide use; whereas, within the remainder of the sample, 50 per cent of patients use thiazides. This result provides some evidence that the coefficient for the thiazide variable may be inappropriate for the validation data set.

For further investigation of this result, we ran model (6) by placing the dummy variable for thiazide use in the model, thus specifying separate calibration-refinement parameters for the

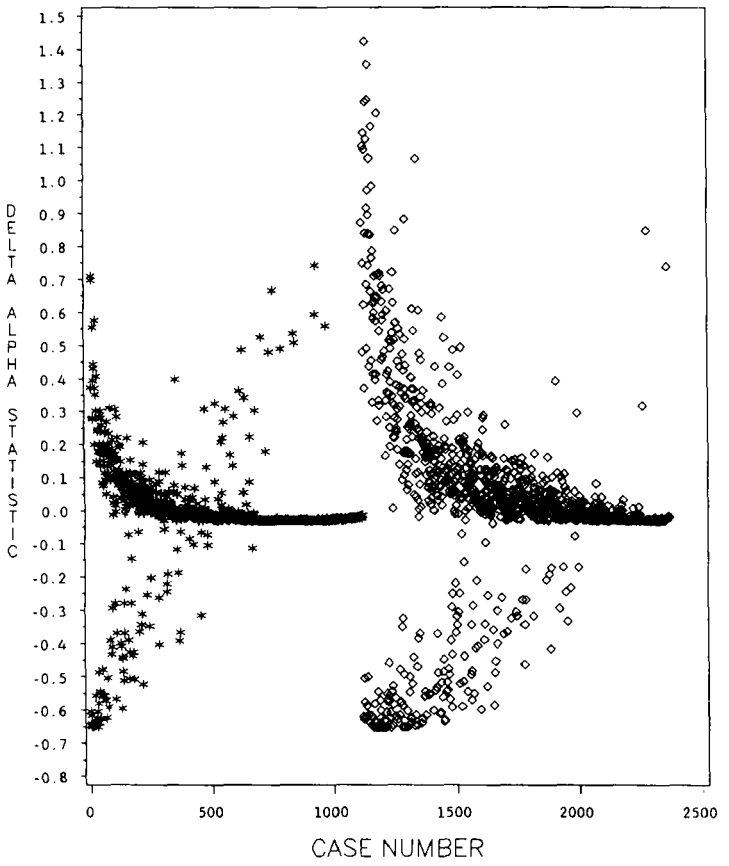

Figure 3. Restructured diagnostic plot for alpha. Thiazide: * no usage; ◇ usage

thiazide group and the non-thiazide group. The predictive equation for this model is

$$\theta_i = -0{\cdot}14 - 0{\cdot}55\,(\text{thiazide}) + 0{\cdot}98L_i - 0{\cdot}09\,(L_i * \text{thiazide}).$$

Based on a likelihood ratio test, we found that the interaction term in this model was not statistically significant; consequently, we dropped it from this reliability model. The subsequent model provided the following reliability equations for the thiazide and non-thiazide users:

$$\text{thiazide:} \quad \theta_i = -0{\cdot}66 + 0{\cdot}93L_i,$$
$$\text{non-thiazide:} \quad \theta_i = -0{\cdot}22 + 0{\cdot}93L_i.$$

Because the calibration coefficient for the thiazide group is basically equivalent to minus the coefficient for thiazide in Table I, it appears that thiazide use is no longer an important predictor of hypokalaemia in our clinic.

Based on the above results, we set the coefficient for thiazide to zero in the original predictive model reported in Table I. We refer to this second predictive model as PM2. Application of Cox's model (3) to the predicted probabilities obtained with use of PM2 resulted in $\hat{\alpha} = -0{\cdot}07$ and $\hat{\beta} = 0{\cdot}94$. For PM2, none of the reliability hypotheses 1–3 specified in Section 2.4 was rejected with use of likelihood ratio tests, thus indicating good calibration and refinement for this model.

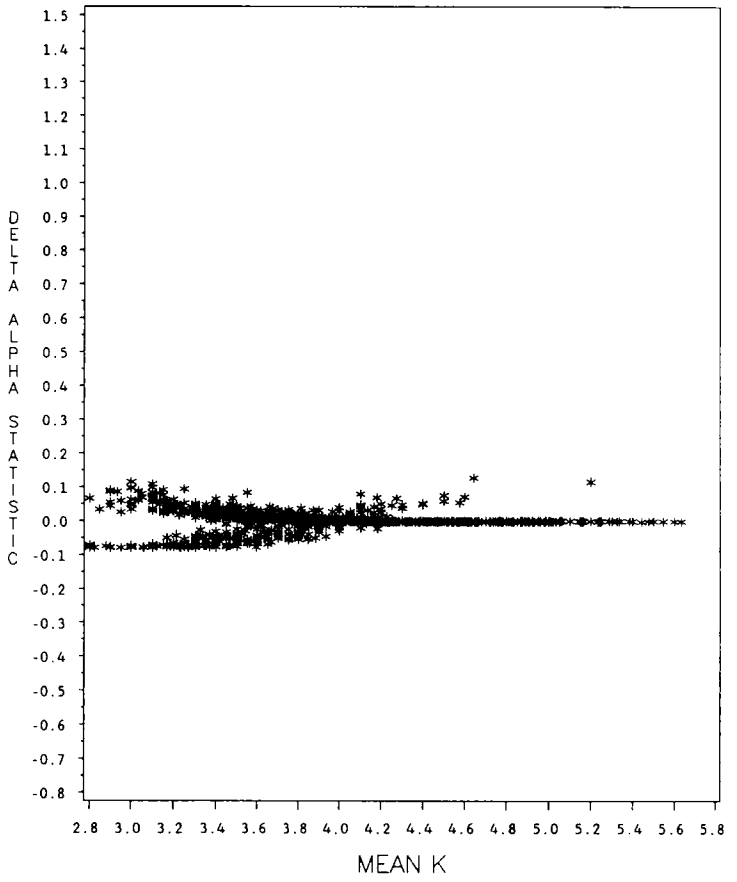

Figure 4. Diagnostic plot for alpha after thiazide correction

The Hosmer–Lemeshow statistic was 15·68 on 10 d.f., also indicating an acceptable fit for the revised model.

Figures 4 and 5 are two diagnostic plots for calibration and refinement obtained after dropping the thiazide variable from the predictive model. These plots provide a striking contrast to those contained in Figures 1 and 2, and further illustrate how the simple correction for thiazide has greatly reduced the influence on calibration and refinement attributable to individual observations. Although the overall tests of reliability indicate that PM2 is an acceptable model, Figures 4 and 5 also provide some evidence that the model still behaves poorly for patients with low values of mean K^+.

Thus, as a final step in this validation exercise, we defined a dummy variable that identified the extreme lower 5 per cent of observations for mean K^+ (mean $K^+ < 3{\cdot}4$), and included this variable in a reliability model of form (6). The results from this model appear in Table III. The estimated refinement parameter within this subgroup is effectively zero ($0{\cdot}06 = 1{\cdot}0 - 0{\cdot}94$), thus indicating that the other predictor variables contribute little additional information for predicting hypokalaemia within this subpopulation. In fact, for a patient who on average is hypokalaemic prior to the current visit, the validation exercise indicates that we can disregard the other variables and assign this patient a 40 per cent probability of hypokalaemia.

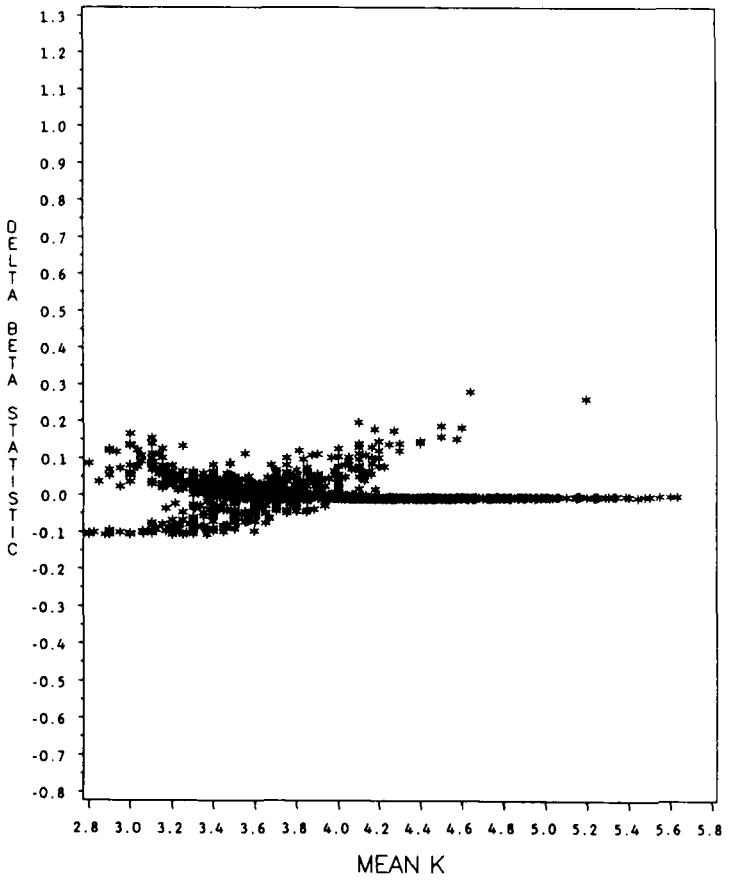

Figure 5. Diagnostic plot for beta after thiazide correction

Table III. Results for interaction model involving low 5 per cent of mean K^+

Parameter	Parameter estimate	Wald statistic	p-value
α_1	0·04	0·06	0·81
Low 5% mean K^+ (α_2)	− 0·43	2·37	0·12
β_1	1·00	144·13	< 0·01
Interaction (β_2)	− 0·94	4·74	0·03

In summary, the results from this validation exercise indicate the need for some minor changes to the original model before we can make valid current applications of the model. Of note, we should remove the thiazide variable as it no longer provides predictive information. This is not surprising since medical practices have changed since 1982. In earlier years, the beginning and maximum doses of hydrochlorothiazide (by far the most common thiazide prescribed) were twice what they are today, and resulted in twice the urinary loss of potassium. Additionally, the validation exercise has indicated that PM1 is not an appropriate model for individuals with a

chronic history of hypokalaemia (mean $K^+ < 3{\cdot}4$). We would have been unable to detect this small subset of patients had we not used the diagnostic procedures proposed here.

5. DISCUSSION

The diagnostic procedures developed within this paper should have use in a wide variety of applications that compare probabilistic determinations and binary outcomes, such as weather predictions, a physician's assessment of the probability of a patient's outcome, and probabilistic predictions based on statistical models. Using a predictive logistic model that contains both discrete and continuous covariates, we have illustrated that use of Cox's model to assess calibration and refinement can lead to an enhanced understanding of the appropriateness of probabilistic predictions. The advantages of this approach are twofold: (a) as Bloch has shown, Cox's approach to test reliability is superior to either the Brier score or Shapiro's logarithmic scoring rule, and (b) Cox's model-based approach permits the application of diagnostic procedures for detection of either individual observations or subsets of the sample that contribute to poor refinement or calibration.

These procedures for external model validation may have important implications for a wide variety of research applications. For instance, with the introduction of quality assurance in health care delivery, there is a proliferation of 'outcomes research' that compares patient outcomes for similar conditions among many health care delivery facilities. To account for the different patient mix in the various facilities, the models developed often account for severity of disease at baseline. If the model developed at one site does not apply at other sites, then these facilities may receive a rating better or worse than they deserve. Therefore, we believe that use of such models at facilities other than at those where they were developed should initially involve the careful application of validation techniques to identify the specific areas of inconsistency between predictions and outcomes. Another problem with such uses of models is that with rapid changes in clinical practices over time, any predictive model for patient outcome may have a limited 'shelf life'. With potential use of the model over a lengthy period, one should conduct routine validation at regular intervals to ensure that conditions in the validation population have not changed.

The goals of our procedures are to identify and possibly correct 'gross' areas of model inadequacy. Just as the development of a model is subject to type I and type II errors, the application of a model to a validation data set is also prone to similar errors. It is therefore important to resist the temptation to use a validation data set to fine-tune a predictive model. The procedures we have proposed here should be used with caution and only in appropriate situations.

ACKNOWLEDGEMENTS

The work of M. E. Miller and W. M. Tierney was supported by grant HS05626 from the Agency for Health Care Policy and Research. The work of S. L. Hui was partially supported by grants AG04518 and AG00313 from the National Institute on Aging.

REFERENCES

1. Cox, D. R. 'Two further applications of a model for binary regression', *Biometrika*, **45**, 562–565 (1958).
2. Harrell, F. E. Jr., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
3. Efron, B. 'How biased is the apparent error rate of a prediction rule?' *Journal of the American Statistical Association*, **81**, 461–470 (1986).

4. Van Houwelingen, J. C. and Le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **9**, 1303–1325 (1990).
5. Miller, M. E. 'Measuring the calibration and refinement of probabilistic determinations', Indiana University School of Medicine, Division of Biostatistics, unpublished technical report 89-1, 1989.
6. Bloch, D. A. 'Evaluating predictions of events with binary outcomes: an appraisal of the Brier score and some of its close relatives', Division of Biostatistics, Stanford University, unpublished technical report 135, 1990.
7. Harrell, F. E. Jr and Lee, K. L. 'Using logistic model calibration to assess the quality of probability predictions', submitted (1990).
8. Lemeshow, S. and Hosmer, D. W. Jr. 'A review of goodness of fit statistics for use in the development of logistic regression models', *American Journal of Epidemiology*, **115**, 92–106 (1982).
9. Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*, Wiley, New York, 1989.
10. Tate, R. F. 'Correlation between a discrete and a continuous variable. Point-biserial correlation', *Annals of Mathematical Statistics*, **25**, 603–607 (1954).
11. Harrell, F. E. Jr., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. 'Evaluating the yield of medical tests', *Journal of the American Medical Association*, **247**, 2543–2546 (1982).
12. Hanley, J. A. and McNeil, B. J. 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **143**, 29–36 (1982).
13. Hanley, J. A. and McNeil, B. J. 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases', *Radiology*, **148**, 839–843 (1983).
14. McClish, D. K. 'Comparing the areas under more than two independent ROC curves', *Medical Decision Making*, **7**, 149–155 (1987).
15. DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. 'Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach', *Biometrics*, **44**, 837–845 (1988).
16. Wieand, S., Gail, M. H., James, B. R. and James, K. L. 'A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data', *Biometrika*, **76**, 585–592 (1988).
17. Brier, G. W. 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, **78**, 1–3 (1950).
18. Shapiro, A. R. 'The evaluation of clinical predictions. A method and initial application', *New England Journal of Medicine*, **296**, 1509–1514 (1977).
19. Pregibon, D. 'Logistic regression diagnostics', *Annals of Statistics*, **9**, 705–724 (1981).
20. McCullagh, P. and Nelder, J. A. *Generalized Linear Models* (2nd edn), Chapman and Hall, New York, 1989.
21. Cook, J. R. and Weisberg, S. *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
22. Tierney, W. M., McDonald, C. J. and McCabe, G. P. 'Serum potassium testing in diuretic-treated outpatients: a multivariate approach', *Medical Decision Making*, **5**, 89–104 (1985).
23. McDonald, C. J., Blevins, L., Tierney, W. M. and Martin, D. K. 'The Regenstrief medical records', *M.D. Computing*, **5**, 34–47 (1988).