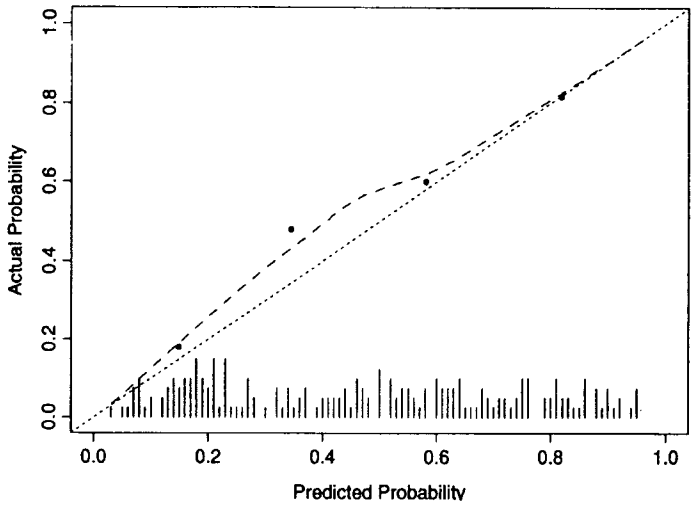

Figure 1. Smooth non-parametric calibration curve (dashed line), subgroup estimates (dots), and ideal relationship (dotted line). The distribution of predicted probabilities is shown above the x-axis. 'Actual probability' is an unbiased estimate of the true probability of response given the level of the predicted probability

When Y is binary and $\hat{Y}$ is the predicted probability that $Y = 1$ versus $Y = 0$, the Brier score[38] or average $(Y - \hat{Y})^2$ is a commonly used mean squared error-type measure of predictive accuracy.

For survival models, one may choose one or more times $(t_1, t_2, \ldots, t_k)$, and plot the predicted probability of surviving until each t_j versus the actual fraction of patients surviving past t_j. The problem here is that we cannot define $Y_i = 1$ if patient i survives past time t_j and then plot the mean Y (by deciles of $\hat{Y}$ or using a smoother) against the mean $\hat{Y}$, since subjects not followed until time t_j are censored, that is, their final outcome status is unknown. One solution is to divide the sample into intervals of $\hat{Y}$ so that there are 50 subjects in each interval of predicted survival, and then plot the mean $\hat{Y}$ within each interval versus the Kaplan–Meier[39] survival estimate at time t_j.

5.4. Shrinkage

Shrinkage is the flattening of the plot of (predicted, observed) away from the 45° line, caused by overfitting. It is a concept related to regression to the mean. One can estimate the amount of shrinkage present (using external validation) or the amount likely to be present (using bootstrapping, cross-validation or simple heuristics). A shrinkage coefficient can be used to quantify overfitting or one can go a step further and use the coefficient to re-calibrate the model. Shrinkage can be defined as a multiplier γ of $X\hat{\beta}$ (excluding intercept(s)) needed to make $\gamma X\hat{\beta}$ perfectly calibrated for future data. The heuristic shrinkage estimator of van Houwelingen and le Cessie[34] (see also reference 40) is

$$\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2}, \tag{2}$$

where p is the number of regression parameters (here excluding any intercept(s) but including all non-linear and interaction effects) and the model χ^2 is the total likelihood ratio χ^2 statistic (computed using the full set of p parameters) for testing whether any predictors are associated

with Y.[§] For linear regression, van Houwelingen and le Cessie's heuristic shrinkage estimate reduces to the ratio of the adjusted R^2 to the ordinary R^2 (derivable from reference 34, Eq. 70).

As an example, suppose that an analyst has considered 10 predictor variables, 6 of which were allowed to enter the model non-linearly (with 2 non-linear terms for each), and tested 8 interaction terms, for a total of 30 degrees of freedom. The model χ^2 is 100 for the full model fit with $p = 30$ d.f. The expected shrinkage is 0·70, indicating that about 0·3 of the model fit is 'noise'. The 'final model' obtained from forward variable selection contains only 3 significant coefficients and has $\chi^2 = 81$, but overfitting is quantified using the 30 candidate d.f. In this example, the number of variables, transformations, and interactions tried was too many for the sample size, and the resulting model is expected to be unstable. As a rough estimate, 0·3 of what was learned from developing the model was really non-replicable noise.

For mild overfitting in the case where the model is needed only to rank likely outcomes and not predict absolute risks, shrinking the regression coefficients will not help since it will not increase real discrimination. If the model is badly overfitted, the model may actually have negative (worse than random) discrimination on new data, and it will have poor calibration. The following heuristic strategy can then be used to determine whether data reduction is likely to result in a model that has any discrimination and how much reduction is required to yield reliable non-shrunken predictions.

First, fit a full model with all candidate variables, non-linear terms, and hypothesized interactions. Let p denote the number of parameters in this model, aside from any intercept(s). Let LR denote the likelihood ratio χ^2 for this full model. The estimated shrinkage is $(\text{LR} - p)/\text{LR}$. If this falls below 0·85, for example, we may be concerned. Let q denote the regression degrees of freedom for a reduced model. In a 'best case', the variables removed to arrive at the reduced model would have no association with Y. The expected value of the χ^2 statistic for testing those variables would then be $p - q$. The shrinkage for the reduced model is then on average $[\text{LR} - (p - q) - q]/[\text{LR} - (p - q)]$. Solving for q gives $q \leqslant (\text{LR} - p)/9$. Therefore, reduction of dimensionality down to q degrees of freedom would be expected to achieve < 10 per cent shrinkage. With these assumptions, there is no hope that a reduced model would have acceptable calibration unless $\text{LR} > p + 9$. If the information explained by the omitted variables is less than one would expect by chance (for example, their total χ^2 is extremely small), a reduced model could still be beneficial, as long as the conservative bound $(\text{LR} - q)/\text{LR} \geqslant 0{\cdot}9$ or $q \leqslant \text{LR}/10$ were achieved. This conservative bound assumes that no χ^2 is lost by the reduction, that is, that the final model $\chi^2 \approx \text{LR}$. This is unlikely in practice, since the data reduction *must* be only X-driven.

As an example, suppose that a binary logistic model is being developed from a sample containing 45 events on 150 subjects. A 10:1 events: d.f. rule suggests we can analyse 4·5 degrees of freedom. The analyst wishes to analyse age, sex, and 10 other variables. It is not known whether interaction between age and sex exists, and whether age is linear. A restricted cubic spline is fitted with 4 knots (requiring two non-linear terms), and a linear interaction is allowed between age and sex. These two variables then need $3 + 1 + 1 = 5$ degrees of freedom. The other 10 variables are assumed to be linear and to not interact with themselves or age and sex. There is a total of 15 d.f. The full model with 15 d.f. has $\text{LR} = 50$. Expected shrinkage from this model is

[§] When stepwise fitting is done, the definition of p is confusing. Many analysts act as if the final model chosen with stepwise variable selection was pre-specified, whether interpreting R^2, confidence limits, or P-values. For estimating the likely shrinkage, it has been shown that p is much closer to the number of candidate d.f. than to the number of parameters fitted in a 'final' model.[40] On a similar note, reference 18 showed how to adjust a linear test of association for having done a test of quadratic effect, concluding that testing the single d.f. statistic for association as if it had 2 d.f. is nearly optimal.

$(50 - 15)/50 = 0{\cdot}7$. Since LR $> 15 + 9 = 24$, some reduction *might* yield a better validating model. Reduction to $q = (50 - 15)/9 \approx 4$ d.f. would be necessary, assuming the reduced LR is about $50 - (15 - 4) = 39$. In this case the 10:1 rule yields about the same value for q. The analyst may be forced to assume that age is linear, modelling 3 d.f. for age and sex. The other 10 variables would have to be reduced to a single variable using principal components or another scaling technique. This single variable may not be interpretable, but using a single score is better than deleting all 10 variables from consideration. If the goal of the analysis is to make a series of hypothesis tests (adjusting P-values for multiple comparisons) instead of to predict future responses, the full model would have to be used.

Bootstrapping[34] and cross-validation[41] may also be used to estimate shrinkage factors. As mentioned above, shrinkage estimates are useful in their own right for quantifying overfitting, and they are also useful for 'tilting' the predictions so that the (predicted, observed) plot does follow the 45° line, by multiplying all of the regression coefficients by $\hat{\gamma}$. However, for the latter use it is better to follow a more rigorous approach such as penalized maximum likelihood estimation,[19] which allows the analyst to shrink different parts (for example, non-linear terms or interactions) of the equation more than other parts.[42]

5.5. General discrimination index

Discrimination can be defined more uniquely than calibration. It can be quantified with a measure of correlation without requiring the formation of subgroups or requiring smoothing.

When dealing with binary dependent variables or continuous dependent variables that may be censored when some patients have not suffered the event of interest, the usual mean squared error-type measures do not apply. A c (for *concordance*) index[1] is a widely applicable measure of predictive discrimination – one that applies to ordinary continuous outcomes, dichotomous diagnostic outcomes, ordinal outcomes, and censored time until event response variables. This index of predictive discrimination is related to a rank correlation between predicted and observed outcomes. It is a modification of the Kendall–Goodman–Kruskal–Somers type rank correlation index[43] and was motivated by a modification of Kendall's τ by Brown *et al.*[44] and Schemper.[45]

The c index is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. The c index measures predictive information derived from a set of predictor variables in a model. In predicting the time until death, c is calculated by considering all possible pairs of patients, at least one of whom has died. If the predicted survival time is larger for the patient who lived longer, the predictions for that pair are said to be concordant with the outcomes. If one patient died and the other is known to have survived at least to the survival time of the first, the second patient is assumed to outlive the first. When predicted survivals are identical for a patient pair, $\frac{1}{2}$ rather than 1 is added to the count of concordant pairs in the numerator of c. In this case, one is still added to the denominator of c (such patient pairs are still considered usable). A patient pair is unusable if both patients died at the same time, or if one died and the other is still alive but has not been followed long enough to determine whether she will outlive the one who died.

Instead of using the predicted survival time to calculate c, the predicted probability of surviving until any fixed time point can be used equivalently, as long as the two estimates are one-to-one functions of each other. This holds for example if the proportional hazards assumption is satisfied.

For predicting binary outcomes such as the presence of disease, c reduces to the proportion of all pairs of patients, one with and one without the disease, in which the patient having the disease had the higher predicted probability of disease. As before, pairs of patients having the same

predicted probability get $\frac{1}{2}$ added to the numerator. The denominator is the number of patients with disease multiplied by the number without disease. In this binary outcome case, c is essentially the Wilcoxon–Mann–Whitney statistic for comparing predictions in the two outcome groups, and it is identical to the area under a receiver operating characteristic (ROC) curve.[46,47] Liu and Dyer[48] advocate the use of rank association measures such as c in quantifying the impact of risk factors in epidemiologic studies.

The c index estimates the probability of concordance between predicted and observed responses. A value of 0·5 indicates no predictive discrimination and a value of 1·0 indicates perfect separation of patients with different outcomes. For those who prefer instead a rank correlation coefficient ranging from -1 to $+1$ with 0 indicating no correlation, Somers' D rank correlation index is derived by calculating $2(c - 0{\cdot}5)$. Either c or the rank correlation index can be used to quantify the predictive discrimination of any quantitative predictive method, whether the response is continuous, ordinal, or binary.

Even though rank indexes such as c are widely applicable and easily interpretable, they are not sensitive for detecting small differences in discrimination ability between two models. This is due to the fact that a rank method considers the (prediction, outcome) pairs (0·01, 0), (0·9, 1) as no more concordant than the pairs (0·05, 0), (0·8, 1). A more sensitive likelihood-ratio χ^2-based statistic that reduces to R^2 in the linear regression case may be substituted.[49–51] Korn and Simon[52] have a very nice discussion of various indexes of accuracy for survival models.

6. MODEL VALIDATION METHODS

As mentioned before, examination of the *apparent* accuracy of a multivariable model using the training dataset is not very useful. The most stringent test of a model (and of the entire data collection system) is an external validation – the application of the 'frozen' model to a new population. It is often the case that the failure of a model to validate externally could have been predicted from an honest (unbiased) 'internal' validation. In other words, it is likely that many clinical models which failed to validate would have been found to fail on another series of subjects from the original source, because overfitting is such a common problem. The principal methods for obtaining nearly unbiased internal assessments of accuracy are *data-splitting*,[53] *cross-validation*[54] and *bootstrapping*.[54–58] In data-splitting, a random portion, for example $\frac{2}{3}$, of the sample is used for all model development (data transformations, stepwise variable selection, testing interactions, estimating regression coefficients, etc.). That model is 'frozen' and applied to the remaining sample for computing calibration statistics, c, etc. The size of the validation sample must be such that the relationship between predicted and observed outcomes can be estimated with good accuracy, and the remaining data are used as the training (model development) sample. Data-splitting is simple, because all the modelling steps, which may include subjective judgements, are only done once. Data-splitting also has an advantage when it is feasible to make the single split with respect to geographical location or time, resulting in a more stringent validation that demonstrates generalizability. However, in addition to severe difficulties listed below, data splitting does not validate the final model, if one desires to recombine the training and test data to derive a model for others to use.

Cross-validation is repeated data-splitting. To obtain accurate estimates using cross-validation, more than 200 models may need to be developed and tested,[54] with results averaged over the 200 repetitions. For example, in a sample of size $n = 1000$, the modelling process (all components of it!) could be done 400 times, leaving out a random 50 subjects each time and developing the model on the 950 remaining subjects. The benefits of cross-validation over data-splitting are

clear; the size of the training samples can be much larger, so less data are discarded from the estimation process. Secondly, cross-validation reduces variability by not relying on a single sample split.

Efron has shown that cross-validation is relatively inefficient due to high variation of accuracy estimates when the entire validation process is repeated.[54] Data-splitting is far worse; the indexes of accuracy will vary greatly with different splits. Bootstrapping is an alternative method of internal validation that involves taking a large number of samples with replacement from the original sample. Bootstrapping provides nearly unbiased estimates of predictive accuracy that are of relatively low variance, and fewer model fits are required than cross-validation. Bootstrapping has an additional advantage that the entire dataset is used for model development. As others have shown, data are too precious to waste.[59,60]

Suppose that we wish to estimate the expected value (for new patient samples similar to the derivation sample) of the Somers' D rank correlation coefficient between predicted and observed survival time. The following steps can be used (see references 55, 58 and 60 for the basic method when applied to binary outcomes):

1. Develop the model using all n subjects and whatever stepwise testing is deemed necessary. Let D_{app} denote the *apparent* D from this model, i.e., the rank correlation computed on the same sample used to derive the fit.
2. Generate a sample of size n with replacement from the original sample (for both predictors and the response).
3. Fit the full or possibly stepwise model, using the same stopping rule as was used to derive D_{app}.
4. Compute the apparent D for this model on the bootstrap sample with replacement. Call it D_{boot}.
5. 'Freeze' this reduced model, and evaluate its performance on the original dataset. Let D_{orig} denote the D.
6. The optimism in the fit from the bootstrap sample is $D_{boot} - D_{orig}$.
7. Repeat steps 2 to 6 100–200 times.
8. Average the optimism estimates to arrive at O.
9. The bootstrap corrected performance of the original stepwise model is $D_{app} - O$. This difference is a nearly unbiased estimate of the *expected value* of the external predictive discrimination of the process which generated D_{app}. In other words, $D_{app} - O$ is an *honest* estimate of *internal* validity, penalizing for overfitting.

As an example, suppose we want to validate a stepwise Cox model developed from, say, a sample of size $n = 300$ with 30 events. The candidate regressors are age, age^2, sex, mean arterial blood pressure (MBP), and a non-linear interaction between age and sex with the terms age × sex and age^2 × sex. MBP is assumed to be linear and additive. Denote these variables by the numbers 1–6. The model χ^2 is 45 with 6 d.f., so the approximate expected shrinkage is $\frac{45-6}{45} = 0{\cdot}87$, or 0·13 overfitting, so some caution needs to be exercised in using the estimated model coefficients and hence in using extreme predicted survival probabilities without calibration (shrinkage). The D for the full model is 0·42. A step-down variable selection using Akaike's information criterion (AIC)[34,61] as a stopping rule (χ^2 for set of variables tested $> 2 \times$ d.f.) resulted in a model with the variables age, age^2, sex, age × sex. The reduced model had $D = 0{\cdot}39$, a typical loss due to deleting marginally important but statistically insignificant variables. Two-hundred bootstrap repetitions are done, repeating the variable selection for each sample using the same stopping rule. We want to detect whether the $D = 0{\cdot}39$ is likely to validate in a new series of subjects from the same population. The first five samples might yield the results shown in Table I.

Table I. Example validation of predictive discrimination

Re-sample	D_{boot} Full model	Variables retained	D_{boot} Reduced model	D_{orig}	Optimism
1	0·45	1,2,3,5,6	0·44	0·37	0·07
2	0·46	1,2	0·34	0·30	0·04
3	0·42	1,2,3,4	0·37	0·34	0·03
4	0·43	1,2,3,5	0·42	0·39	0·03
5	0·41	1,3,4	0·39	0·37	0·02

The average optimism is 0·038, so the bootstrap estimate of the expected validation of D_{app} is $0·39 - 0·038 = 0·352$. The analyst may or may not be worried about the 0·038 overfitting, but the best estimate of predictive discrimination is $D = 0·352$ – this is a better estimate of the likely 'external' validation accuracy than is 0·39 if all other aspects of the study design remain constant. The $D = 0·352$ is the honest estimate of predictive accuracy that should be quoted when the researchers document the accuracy of the reduced model that was developed on the entire dataset using a stepwise variable selection algorithm.

It is usually informative to repeat the bootstrap validation with and without stepwise variable selection. Usually, the amount of predictive information lost by deleting marginal variables is not offset by the decreased optimism of the stepwise model. One way to demonstrate this point is to observe how often 'insignificant' clinical predictors have clinically sensible signs on their regression coefficients. Stepwise variable selection, which requires binary decisions about the inclusion of variables (unlike shrinkage), causes information to be lost.[2]

The same strategy can be used to estimate the over-optimism in an R^2 measure[49] from the original model fit. For estimating the prediction error at time t in a survival model, similar steps could also be used. Instead of validating a correlation D, we substitute for example the statistic D = difference between mean predicted 2-year survival probability and Kaplan–Meier 2–year survival estimate. The survival estimates are made by, say, deciles of predicted 2-year survival from the original model fit using the following steps, for example:

1. Develop the model using all subjects.
2. Compute cut points on predicted survival at 2 years so that there are m patients within each interval ($m = 50$ or 100 typically).
3. For each interval of predicted probability, compute the mean predicted 2-year survival and the Kaplan–Meier 2-year survival estimate for the group.
4. Save the apparent errors – the differences between mean predicted and Kaplan–Meier survival.
5. Generate a sample with replacement from the original sample.
6. Fit the full model.
7. Do variable selection and fit the reduced model.
8. Predict 2-year survival probability for each subject in the bootstrap sample.
9. Stratify predictions into intervals using the previously chosen cut points.
10. Compute Kaplan–Meier survival at 2 years for each interval.
11. Compute the difference between the mean predicted survival within each interval and the Kaplan–Meier estimate for the interval.
12. Predict 2-year survival probability for each subject in the original sample using the model developed on the sample with replacement.

13. For the same cut points used before, compute the difference in the mean predicted 2-year survival and the corresponding Kaplan–Meier estimates for each group in the original sample.
14. Compute the differences in the differences between the bootstrap sample and the original sample.
15. Repeat steps 5 to 14 100–200 times.
16. Average the 'double differences' computed in step 14 over the 100–200 bootstrap samples. These are the estimates of over-optimism in the apparent error estimates.
17. Add these over-optimism estimates to the apparent errors in the original sample to obtain bias-corrected estimates of predicted versus observed, that is, to obtain a bias- or overfitting-corrected calibration curve.

7. SUMMARY OF MODELLING STRATEGY

1. Assemble accurate, pertinent data and as large a sample as possible. For survival time data, follow-up must be sufficient to capture enough events as well as the clinically meaningful phases if dealing with a chronic disease.
2. Formulate focused clinical hypotheses that lead to specification of relevant candidate predictors, the form of expected relationships, and possible interactions.
3. Discard observations having missing Y after characterizing whether they are missing at random.‖ See reference 62 for a study of imputation of Y when it is not missing at random.
4. If there are any missing Xs, analyse factors associated with missingness. If the fraction of observations that would be excluded due to missing values is very small, or one of the variables that is sometimes missing is of overriding importance, exclude observations with missing values¶. Otherwise impute missing Xs using individual predictive models that take into account the reasons for missing, to the extent possible.
5. If the number of terms fitted *or* tested in the modelling process (counting non-linear and cross-product terms) is too large in comparison with the number of outcomes in the sample, use data reduction (ignoring Y)[20–23] until the number of remaining free variables needing regression coefficients is tolerable. Assessment of likely shrinkage (overfitting) can be useful in deciding how much data reduction is adequate. Alternatively, build shrinkage into the initial model fitting.[19]
6. Use the entire sample in the model development as data are too precious to waste. If steps listed below are too difficult to repeat for each bootstrap or cross-validation sample, hold out test data from all model development steps which follow.
7. Check linearity assumptions and make transformations in Xs as needed.
8. Check additivity assumptions and add clinically motivated interaction terms.
9. Check to see if there are overly-influential observations.[30] Such observations may indicate overfitting, the need for truncating the range of highly skewed variables or making other pre-fitting transformations, or the presence of data errors.

‖ For survival time data, no observations should be missing on Y. They should only have curtailed follow-up.

¶ Alternatively, impute missing values for the predictor but perform secondary analyses later to estimate the strength of association between X and Y after deleting observations with that predictor imputed, as imputation will attenuate the relationship.