Received: 6 January 2010, Revised: 21 February 2010, Accepted: 23 February 2010, Published online in Wiley InterScience: 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1311

Deterministic fallacies and model validation

Douglas M. Hawkins[a*] and Jessica Kraker[b]

Stochastic settings differ from deterministic ones in many subtle ways, making it easy to slip into errors through applying deterministic thinking inappropriately. We suspect this is the cause of much of the disagreement about model validation. A further technical issue is a common misapplication of cross-validation, in which it is applied only partially, leading to incorrect results. Statistical theory and empirical investigation verify the efficacy of cross-validation when it is applied correctly. In settings where data are relatively scarce, cross-validation is attractive in that it makes the maximum possible use of all available information, at the cost of potentially substantial computation. The bootstrap is another method that makes full use of all available data for both model fitting and model validation, at a cost of substantially increased computation, and it shares many of the broad philosophical background of cross-validation. Increasingly, the computational cost of these methods is not a major concern, leading to the recommendation, in most circumstances, to use cross-validation or bootstrapping rather than the earlier standard method of splitting the available data into a learning and a testing portion. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: bootstrap; diagnostics; model validation; cross-validation; stochastic

1. INTRODUCTION

The distinction between deterministic and stochastic settings is often overlooked, and it is easy to slip into traps of applying deterministic thinking to stochastic problems. Most familiarly, perhaps, early in our school years we learned the algebraic law of transitivity. That is, if X, Y, and Z are some mathematical constructs such as constants, vectors or functions and $X = Y$ and $Y = Z$, then necessarily $X = Z$. In our subsequent statistics classes, we may have though this same law applied in settings such as

> If A is not statistically significantly different from B, and B is not statistically significantly different from C, then presumably A and C are also not statistically significantly different.

Our teachers soon disabused us of this notion, telling us that it was possible, and indeed not uncommon, for A and C to be enormously significantly different, even if each was statistically indistinguishable from B.

While every graduate of an introductory statistics class has absorbed this lesson, more subtle variants of this deterministic misconception can persist. One, perhaps, lies within the concept of 'the' predictive model. It may be convenient to think of a model as something fixed and immutable, like say a chemical property of a molecule, but it is not. It involves a mathematical structure, and a number of parameters whose values have to be estimated from data. The description of, say, PLS regression explaining a property of a chemical from a particular set of molecular descriptors hides the fact that the PLS regression requires a vector of coefficients, and the values of these coefficients change depending on exactly what data are used to estimate them. The discussion about validation is intended to apply to predictive models in general. The concepts will be phrased using QSAR modeling language, but this should not be taken to mean that QSAR is the only area to which they apply.

It will be convenient to use a shorthand characterization of a predictive model. We will use the linear model for this purpose: given an activity **y** and a set of structure measures **X** (containing the various predictors as its columns), a linear model would hold that

$$y = \mathbf{X}\beta + e \qquad (1)$$

where the parameters are the vector $\boldsymbol{\beta}$ (and an intercept is also needed if we choose not to pre-center the data). This shorthand is not restrictive; our actual model could be linear or nonlinear; it could be a neural net; a support vector machine; or nearest neighbors. Each of these methodologies involves some tuning using data, and we will use the phrase 'estimate $\boldsymbol{\beta}$' as shorthand for whatever is involved in actualizing the model. This might involve deciding how many components to use in a PLS, followed by estimation of the resulting coefficients; or it might be deciding how many neighbors to use in a nearest neighbor prediction.

In order to assess the utility of the model, an estimated model must then be validated. What do we mean by 'model validation?' Operationally, this means that the elements of the 'error vector' **e** must be

(1) acceptably small, so that the model is making acceptably accurate predictions,
(2) unrelated to the predictors in **X**, otherwise the model could be improved by somehow incorporating the not-yet-modeled dependency of **e** on **X**, which means that the current model is inferior to this refinement,
(3)

* *Correspondence to: D. M. Hawkins, School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN 554, USA.*
E-mail: doug@stat.umn.edu

a *D. M. Hawkins*
School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN 554, USA

b *J. Kraker*
Department of Mathematics, University of Wisconsin, Eau Claire, WI, USA

unrelated to other factors that are known about the compounds but not necessarily captured in **X**. For example, if a compound library contains alkanes and alkenes but the predictors used in **X** do not differentiate between the two classes, we should check separately whether the **e** of alkanes and alkenes meet the first two requirements.

Expanding on this third point, suppose we have two data sets. One is the data set that was available at the time of the model fitting, and the other became available only some time later. In that case, we would have to do the fitting using the first, available, data set and prediction for the latter would be the verification that the model fitted in the original framework fits this later data as well as the earlier data. This is, of course, the sort of validation advocated by those who believe that testing must be applied to a brand new data set.

As a preview to the ideas of model *validation*, we should note that once we have statistics involved, there is no such thing as model validation. Model validation consists of proving a null hypothesis, and this can never be done. In conventional statistical testing, accepting a null hypothesis does not prove that it is true; it merely reflects that the attempts to falsify it were unsuccessful. Similarly, model validation consists of checks to see whether the fitted model failed by, for example, producing **e** that did not behave like random values from some statistical distribution. If the model fails to fail, then we declare that it has passed the validation check. But this does not mean that it would not fail if applied using some fresh information. For example, if the model underpredicted alkanes and overpredicted alkenes, but we did not separate out these two classes when looking at model fit, then an initial analysis might show nothing amiss, but a more refined analysis including this distinction would cause it to fail. That is, even if the model does not fail by the validation process, it might well fail a more searching validation assessment—for example one introducing the additional information about the presence of different chemical classes. The validated model can therefore be something quite ephemeral.

2. THE ROLE OF STATISTICAL THINKING

In the prototype Equation (1), the activity **y** is predicted by $\mathbf{X}\boldsymbol{\beta}$ with some 'error' **e**. This 'error' can arise from two conceptually different sources. One is traditional experimental error. If the dependent (or perhaps the predictors) involves an imprecise measurement, then there is the sort of traditional random error that underlies standard statistical discussion and that everyone understands. For example, **y** might be something like a laboratory test of mutagenicity, where the inherent random variability of measurement leads to differences between the model $\mathbf{X}\boldsymbol{\beta}$ and the actually observed **y**.

There is another potential data-related source. It may be that there can be many different compounds with the same **X** vectors, but inherently different activities **y** even if there were no issue of measurement error. In this case, the model necessarily cannot exactly predict the activity of all of them. At best, the model may overpredict some compounds, underpredict other compounds, and be about right on yet other compounds. In symbols, some of the molecules may have large positive **e**, others large negative **e** and still others near-zero **e**.

What is the reason for this discrepancy? As the basis tenet of modeling is that function follows form—that activity is a consequence of molecular structure—we would have to ascribe this imperfection to other properties of the molecule that our collection of descriptors does not capture. Maybe if we could enrich our collection of descriptors we could get rid of these misfits, but with the available descriptors we cannot.

Such a collection of errors **e** is deterministic, not stochastic, yet it may be helpful for us to treat the errors as if they were stochastic (References [1]: p. 59 and [2]). To explore this, suppose we create a gambling game. You will make a list of some of the significant dates in your life—perhaps your date of birth, those of your parents, your graduation date—and write them out as 8 digit numbers. For example if you were born on 15 August 1951, your number would be 19510815. I will pick a transcendental number like e or $\sqrt{\pi}$. We will then bet on whether the 19510815th digit of my number is even or odd. The bet can then be repeated using the other dates. Although perfectly deterministic, you could for all practical purposes treat this game as if the successive outcomes were generated by tossing a fair coin.

This sort of 'deterministic disguised as stochastic' setup perhaps underlies some of the thinking in QSAR with non-zero **e** from this second cause. If there is no measurement random variability then the quantities **e** are deterministic, but provided we look at our compounds in a way that does not systematically favor some over others based on their **e** values, we can treat the **e** as having been randomly sampled from some statistical distribution.

Clearly the realism of this thinking depends on exactly how the **X**, **y** data set is selected and processed. If we look at our compounds in increasing order of molecular weight, then the **e** might have some structure that would make nonsense of the stochastic thinking.

Regardless of this particular issue, there is an inherent assumption that we are interested in the applicability of our model to some larger 'population' of compounds, with the data at hand being reliable representatives of this population—perhaps by plausibly being a random sample from this population. Unless the available data satisfy this requirement of being random-like representatives of this larger population, it is hard to imagine how one can use them to make inferences about a broader class of compounds.

3. THE IMPACT OF SAMPLE SIZE

A predictive model involves two elements. One is deciding what methodology to use—PLS, ANN or k-NN for example. The other consists of estimating the parameters of the models, for example, the number of components and resulting coefficient vector in a PLS. The quality of the resulting predictions is affected by the quality of this second task, even if the underlying methodology and model are valid.

We can easily quantify this dependency for idealized linear models. If the relationship is linear as assumed and **e** follows a normal distribution with constant variance σ^2, then the variance of the prediction of a future compound is

$$\sigma^2(1+h), \qquad (2)$$

where h, the leverage, reflects the random error in the estimate of $\boldsymbol{\beta}$; and the position in **X** space of the compound whose activity is being predicted (Reference [3]: pp. 161 and 177).

If h is small, then this prediction variance is essentially just σ^2, but large values of h make this second term increasingly important. Two features impact the size of h:

- the random variability in the estimate of $\boldsymbol{\beta}$ and
- the location in **X** space where the prediction is being made.

The first issue encourages large calibration samples. Not just in linear modeling, but broadly throughout model fitting, larger samples of valid data lead to lower random variability in fitted parameters. Thus to have better models that make more accurate predictions, we want to include as many valid data cases as possible.

The second enjoins against extrapolation. Loosely speaking, to keep this term small, we want to calibrate using compounds that cover as wide a domain of **X** space as possible. For this too, we should use as many valid data cases as possible for the fitting—more compounds will give us information from more extreme molecules, and will also better fill in the space between molecules.

Both issues involve diminishing returns to scale. If we have a large calibration data set with only a few parameters to estimate, then the h values may be so small they are overwhelmed by the '$1+$' in Equation (2). Then cutting the size of the calibration set in half, say, would cause us little harm. But to the extent that we have relatively few compounds and/or large numbers of parameters to estimate, the h values may be appreciable, and thus a reduction in the calibration sample size has harmful effects.

Sample size has a perhaps even larger impact on the validation exercise. Summary measures of fit often involve residual sums of squares. These have a notoriously variable distribution that, even in the absence of outliers, can be largely determined by a small fraction of the data. Thus if we want to assess the quality of the model using some metric based on a residual sum of squares, we need a large validation sample size to remove as much as we can of the large random variability from the residual sum of squares.

In other words, unless we have simple models with few predictors and many compounds, we want to maximize the size of both the calibration and the validation data sets—see References [4,5].

4. MORE ABOUT SEPARATE CALIBRATION AND VALIDATION DATA SETS

A central tenet of science is that results must be reproducible—other scientists following a researcher's instructions must be able to duplicate the results in their labs. The reason few people currently believe in the Fleischmann–Pons claim of cold fusion is that others following their recipes were unable to achieve the same outcomes they had claimed.

As modeling is a scientific exercise, it should also be subject to the same requirements. Some people interpret this to mean that you must get some fresh compounds, not used in the original modeling exercise, and see if you can predict the activity of these compounds also. We would argue that the reproducibility actually required by the scientific process is much different from this. Our view is that, if you provide your data set and predictors, and describe your criteria and algorithms, other scientists should be able to duplicate your fits. This requirement does not involve any mention of fresh data sets. We do not dispute that getting fresh data is a valuable exercise, but we do disagree on whether fresh data are an essential part of model validation.

But let us consider the setting in which we have one data set that we use for learning and calibration, and another that we can use for testing but that we do not use (whether from choice or necessity) in the calibration.

Several different shades of this scenario come to mind. One is that timing issues are involved. The calibration exercise had to be completed by year-end, but the new validation data only became available in the New Year. We are thus forced to use only the data available at the time of calibration.

Another is that the data set involves experimental measurement, and that we have some data from lab A that we want to use for all the modeling, but will make use of the data from lab B for model testing.

A third is that we have two chemical libraries—for example one of alkanes and another of alkenes. We would like to know whether a single model fits both, and plan to do this by fitting the model to the alkanes and then seeing if it applies also to the alkenes.

In all these settings, it is legitimate—but not necessarily a good idea—to carry out the model fitting using just the first data set for calibration, and then to do some checking by applying this model to the second validation (also known as 'test') data set. In each of these cases, the model validation could have two outcomes:

- the model could pass the check,
- the model could fail the check.

If the model passes the check, then it would immediately become obsolete. This is because, if we are happy that the calibration and validation data are fitted by the same model, then surely the next step would be a combined analysis in which we refitted using a data set comprising both sources. Such a data set, being larger than the original calibration data set, would have a wider domain of applicability. Being based on a larger sample size, it would also have more precise parameter estimates leading to better future predictions. In other words, we would replace the original starting model that we were validating with another, better model.

If the model fails the check, then we would know that the same model did not fit the calibration and the validation data set. We then probably go back to the drawing board to broaden the model to incorporate additional features so that it was able to accommodate both data sets, or to more clearly demarcate the types of data to which the model applied.

In the last two of these three settings, both the data sets are available at the time of the modeling and we could potentially study them in a combined analysis. In the case of linear models, there is a standard methodology for testing whether the same model fits both data sets A and B—see, for example, Chapter 14 in Reference [1]. Create an indicator vector **b** whose elements are 0 for compounds in set A and 1 for compounds in set B. Also create a matrix **W** which copies **X** in rows corresponding to compounds in set B, and has zeros in rows corresponding to compounds in set A. Then fit the model

$$\mathrm{y} = \mathbf{X}\beta + \delta\mathbf{b} + \mathbf{W}\gamma + \mathrm{e} \tag{3}$$

introducing additional parameters δ and (vector) $\boldsymbol{\gamma}$. Provided some linear model fits each of the data sets, this is a powerful tool for checking whether the model is the same one.

Specifically, if the data suggest that $\boldsymbol{\gamma}$ is nonzero, you conclude that data sets A and B do not conform to the same linear model. You should probably separate them out and fit different models. If however, it looks like $\boldsymbol{\gamma}$ is zero, but δ is nonzero, sets A and B conform to the same model overall, but the **y** values of one set are systematically higher, for the same **X**, than the other. Statistically, this is the analysis of covariance model. In this setting, enriching **X** by adding the indicator **b** to it will lead to a single model that accommodates both data sets and that benefits from the synergy of estimating the common parameters $\boldsymbol{\beta}$ from a larger data set.

Finally, if it appears that both δ and $\boldsymbol{\gamma}$ are zero, then conclude that the model is validated, with a common model explaining both the A and the B data sets. You can then combine them into a single data set and refit to get an enhanced estimate of $\boldsymbol{\beta}$.

5. SPLIT SAMPLES AND CROSS-VALIDATION

For many years, the standard practice in statistics was to validate models by randomly splitting the available data into a separate calibration and validation data set. All modeling steps were then performed on the calibration data set and the model assessed by how well the resulting fit explained the validation data set. This approach largely originated in social science applications when researchers started to use multivariate methods whose theoretical properties were poorly understood.

The split sample approach continues to be seen in the chemometric literature, but has largely vanished from the statistical. Its replacement there has been the resampling paradigm—see References [5–8]. In this, the entire sample is used for calibration, and the entire sample is also used for validation. In the light of the earlier comments about the desirability of large samples for both the calibration and the validation parts of the exercise, this maximum use of both is attractive.

There is one qualifier on this point—there is a diminishing return to scale in sample size: if the calibration sample is already large, then making it larger will not make predictions much better. But if the calibration sample is small, making it yet smaller will make predictions much worse. Similar remarks apply to the validation sample. This means, in the present context, that if we have a large sample available, the traditional method of splitting it into a calibration and a validation sample will be acceptable, but if only a modest sample size is available, splitting it into two will seriously degrade both the calibration and the validation activities. In the typical early social science experiment where the split-sample approach became popular, it was easy to get very large samples, leading to separate large calibration and large validation data sets. Thus the calibration exercise would lead to small h values that would not benefit much from larger samples, and the validation exercise would include enough cases to largely remove the effect of random variability on sums of squares performance measures, and so the split sample approach worked very well.

But once we leave this situation of very large sample availability, making maximum use of the sample for both calibration and validation is highly advisable. In particular, *properly applied* cross-validation will yield both better models and better validation of these models than we would get by splitting the available sample into separate calibration and validation samples. Statistical theory predicts this [8], and it has been verified empirically in a number of simulated [5,7] and chemometric exercises [4,9,10]. As huge samples and simple models are the exception rather than the norm in modern chemometric settings, we believe this supports cross-validation as a default validation tool.

There is another argument against taking the available compounds and splitting them into a calibration and a validation sample—this is that both the fitted model and the results of the validation will depend on exactly which compounds were put to each use. In the huge sample setting, the law of large numbers will largely make this concern go away, but in the setting of more modest sample sizes, it remains an issue. It is surely disconcerting that two researchers, using the identical data sets and algorithms, but different valid (for example random) partitions into the calibration and the validation data set could reach substantively different conclusions. But this is the implication of using the split-sample approach in modest-sized data sets.

There is an important assumption in cross-validation. This is that the compounds in the sample are 'exchangeable'. This assumption is violated if the data set contains known structure that is not reflected in the descriptors: for example, if the collection contains compounds of different chemical class but this is not reflected in the descriptors and the mechanism of these classes differ. Note that this is a soluble problem—see the earlier comments on use of indicator variables. The split-sample method, of course, also involves an exchangeability assumption, though perhaps less obviously, and so does not escape the issue.

We statisticians have to trust that our chemical colleagues supplied a set of compounds that is representative of a single population, but recall that, as outlined above, there are tools to check whether identified different classes may be pooled. Otherwise, if we know our collection of compounds is heterogeneous, we can add indicators to our set of descriptors, as outlined in the previous section, and check whether the chemical heterogeneity corresponds to different model needs. By virtue of the added indicator variables, the necessary exchangeability can largely be restored.

It can be noted that cross-validation uses only the compounds available in the sample, and so tautologically does not tell us what will happen in a future sample. This is, of course, equally true of a sample that is split into a calibration and a validation sample. We will only know what happens with future samples when the future samples become available. And then we will still not know what the post-future samples will produce. In other words, as in all science, our understanding evolves over time as new information becomes available. However to make any prediction process worthwhile, we must have reason to believe that the available data set (whether used for calibration, validation or both) is *representative* of the data for which we want to predict outcomes.

There are various 'flavors' of cross-validation available; both Shao [8] and Hastie *et al*. [5] provide excellent summaries of cross-validation techniques. While the described concepts apply to general forms of cross-validation, we would argue for the following precepts:

(1) Leave-one-out, LOO, cross-validation is the most familiar. In this, each compound in turn is omitted and predicted using the remaining $n-1$ compounds. LOO has some advantages of being familiar and exactly reproducible, but is slow for large samples. If used for model selection, it has a tendency to select models that are slightly too big—for example include

one or two extraneous predictors [8]. To avoid this, *holdout samples larger than one are suggested*. These scale to large samples well, and as Shao [8] has shown mathematically, larger holdout samples are better for selecting models of the right size. The standard approach for doing this is *M*-fold cross-validation.

(2) *Monte Carlo or M-fold cross-validation are user-friendly methods*
   - Monte Carlo cross-validation, simply put, randomly splits the data into a calibration set and validation set repeatedly. See Shao [8] for a complete explanation; this method is also utilized below, applied to a nearest-neighbors model.
   - a variety of cross-validation known as *M*-fold cross-validation (see Reference [4]), in which the data set is randomly divided into *M* different groups (typically 5 or 10 [5]). The intended analysis is applied to all data to get the final model. Then *M* separate analyses are performed. In each, the full model fitting procedure is applied to $M-1$ of the groups, and the fit applied to the holdout group to get the out-of-sample predictions and prediction residuals.

(3) *In M-fold cross-validation, repeated splits of the data should be made*. By definition, Monte Carlo cross-validation involves a large number of splits of the data set, in each of which the available data are split randomly into *M* groups to be used for the fitting and testing (see Reference [10]). The criterion (e.g. predicted sum of squares) is averaged over all the repeated splits, so as to not tie the measure to one particular division of the data.

One other feature of cross-validation is crucial. Cross-validation involves carrying out the modeling exercise on the full data set, and then redoing the modeling exercise repeatedly on a series of subsets of the data set. It is essential that these repeats incorporate *all* steps involved in the modeling process. Some methodologies which involve 'feature selection' illustrate this point. Examples are: nearest neighbors in which we need to find how many neighbors (*k*) should be included; PLS in which we need to decide how many components to use; and subset regression in which we decide which features to keep and which to drop. The feature selection is commonly and effectively done by a cross-validation within the cross-validation (a so-called 'double' or 'two-deep' or 'correct' cross-validation [4,5,10,11]).

To illustrate the issue, consider the nearest neighbors' model and the steps involved in the model-fitting process, as applied to a not-held-out (used to calibrate) and a held-out (used to test) data set:

(1) fit the nearest-neighbor model (for several *k*) on the not-held-out data;
(2) use an internal process, say internal cross-validation, to choose the number of neighbors *k* resulting in the best prediction model (according to some criterion, for example minimum predicted sum of squares [10]), of the not-held-out data;
(3) use this model with the selected *k* to predict the held out data and produce an appropriate measure of predictive ability of the model.

This is the process we would follow if we insisted on testing on a brand new data set, but applies also in a cross-validation setting.

Note that the internal cross-validation of step 2 *does not* result in a cross-validated estimate of the predictive ability of the model. In order to get such a measure, we would do this splitting into a not-held-out calibration and held-out validation set repeatedly (external cross-validation). Then *for each such split*, we must apply the internal cross-validation to the not-held-outset in order to select the best *k* for that set; then using that *k*, we predict the held-out set. The predictions from the various splits of the external cross-validation are combined into a cross-validated estimate of the predictive ability of the model.

So in a proper cross-validation assessment, we will determine the appropriate neighborhood size for the full data set to obtain the overall fitted model, and will then need to determine it anew for each of the calibration subsets that comes up in the cross-validation. Each of these folds in the cross-validation could produce a different neighborhood size *k*. These differences in the selected *k* from one fit to another cause confusion in some circles. They are not evidence of a deficiency in cross-validation. The neighborhood size is one of the parameters of the fitted model, and this is determined by the full-sample analysis. The possibly different *k* values obtained in the cross-validation fits are therefore no more disturbing than the differences in say the regression coefficient vectors obtained in the different cross-validation folds.

Similarly, if we are using subset regression, we need to repeat the subset selection for each of the calibration subsamples generated in the cross-validation, and could find different predictor subsets being selected in the different folds. These differences are interesting and may repay study, but are not a commentary on the validation itself except insofar as they indicate the stability or instability of the final model – the model fitted to the full data set.

It is not uncommon to see violations of this central principle in the literature. For example, an investigator may decide from the full sample that using six neighbors seems best, and will then use six neighbors for each of the 'folds' in the cross-validation. This invalid analysis, which we have termed 'naïve cross-validation' [4,10] will lead to incorrect conclusions and must be avoided.

6. CASE DIAGNOSTICS

Once a model has been fitted, we are able to make predictions and calculate residuals. These can be calculated for the compounds that were used in the calibration, but more valuably can be 'predicted' quantities obtained with those compounds excluded from the calibration. It is an attraction of cross-validation that (unlike split samples) it produces predicted values and prediction residuals for each and every compound in the data set. Many methodologies also produce values of the 'leverage', a measure of how unusual the compound's **X** vector is. Conventional statistical regression diagnostics [1–3] include plotting the prediction residuals versus the predicted values, and against the leverage or a function of the leverage. These plots are invaluable in diagnosing departures from model assumptions such as linearity, in identifying outlying cases (those with large prediction residuals) and those with high leverage. These, along with outliers, should be diagnosed carefully to see if there is any explanation for their atypicality.

If the data set contains identifiable subgroups—for example those provided by different investigators—then marking these residual plots by the subgroup will help identify possible problems of the model fitting some subgroups differently than others.

Comparison of the $M+1$ distinct fits generated in an *M*-fold cross-validation can also yield valuable information. Sometimes, in methods that involve feature selection, there are quite large differences in the features selected in the different fits. It has been suggested that this is a failing of the cross-validation methodology. We believe, on the contrary, that this is a strength—it is a valuable indicator of the robustness of the original full-sample fit. For example, if any of these *M* fits of the partial data sets is much different from the full-sample fit, then that is an indication that the corresponding held-out set of compounds is particularly influential in the overall model fit. This is warning you that, had your colleague in Columbus not sent you his particular suite of compounds, or if your literature search had failed to turn up the last ten compounds used in the study, the conclusions of your analysis would have been substantially different. While not strictly germane to the assessment of your model, this is surely a valuable thing to know.

7. CROSS-VALIDATION AND THE BOOTSTRAP

In *M*-fold cross-validation, we repeatedly split our *n* compounds into a calibration set of size $n(M-1)/M$ and a validation set of size nM. We use the model fitted to the first set to predict the second set and summarize the quality of these predictions to assess the model itself. Conventional discussions of cross-validation do not stress the point made above—that much may be learned by looking, not just at the predictions from these different models, but by comparing the fitted models themselves—but this is indeed the case.

The bootstrap comes in many varieties [12]. In the most familiar, we repeatedly take a random sample of size *n* from our *n* available compounds, sampling with replacement (so the same compound can be included more than once), and perform the modeling operations on these samples. The quality of the original model is assessed by summary statistics of the different models fitted. Some discussions of bootstrapping note that, on average, one-third of the compounds are omitted from each of the bootstrap samples, and the model fitted to the bootstrap sample could be used to predict these 'out of bag' compounds, leading to case diagnostics for each non-included compounds in each bootstrap fit. These predictions are conceptually very similar to what one would get from a 3-fold cross-validation—the main difference is that in a 3-fold cross-validation the calibration set includes only one copy of each included compounds while in the bootstrap there may be several copies.

This side-by-side discussion of *M*-fold cross-validation and the bootstrap shows that the two methods have much in common. Both use the entire available data set for their model fitting and for their model assessment. They have comparable computational complexity. While the emphasis in cross-validation is usually on the predictions of the hold-out compounds, and that of the bootstrap is usually on the different fitted models, comparison of the different models in cross-validation (as done in the bootstrap) is a valid and useful exercise, as is examination of the out-of-bag predictions (as done in cross-validation) in the bootstrap.

8. CONCLUSION

The issue of how to validate models has plagued those involved in modeling for many years. Early work in the social sciences addressed this by splitting out the generally quite large available data sets into separate learning and validation data sets, applying then-new methodologies such as multivariate methods to the learning data sets and testing the resulting models using the hold-out validation data set. This method is computationally economical, and effective in settings where sample is plentiful. But this is not the situation in which many modelers find themselves.

Our late colleague Seymour Geisser famously claimed that all statistical problems are really prediction problems under varying levels of disguise, and pioneered the methodology of sample reuse, in which all available sample is used for both learning and validation, at a cost of vastly increased computational load. This thinking has been extended in the statistical literature over the last 30 years and the methods of Stone and Geisser refined and extended. We believe that this approach of sample reuse should be the standard in most modeling, with all available data being used, reserving the old and more wasteful paradigm for settings where very large samples are available.

This involves assumptions of exchangeability, which come from the chemical construction of the data set and are based on the professional knowledge of chemistry colleagues. Thus, we believe the equation (in notation common to both chemistry and statistics) that

$$\text{Good Chemistry} + \text{Good Statistics} = \text{Good Science}$$

and champion closer collaboration between the two professions to ensure appropriate exchange of expertise.

Acknowledgements

The authors are grateful to the reviewers for their helpful suggestions for improvement in the paper.

REFERENCES

1. Draper N, Smith H. *Applied Regression Analysis*, (3rd edn). John Wiley & Sons, Inc: New York, 1998.
2. Cook R, Weisberg S. *Residuals and Influence in Regression*. Chapman and Hall: New York, 1982.
3. Cook R, Weisberg S. *Applied Regression Including Computing and Graphics*. John Wiley & Sons, Inc: New York, 1999.
4. Hawkins D. The problem of overfitting. *J. Chem. Info. Comput. Sci.* 2004; **44**: 1–12.
5. The Elements of Statistical Learning. http://www-stat.stanford.edu/~tibs/ElemStatLearn [4 Jan, 2010].
6. Stone M. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc., Ser. B: Methodol.* 1974; **36**: 111–147.
7. Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. *Int. Stat. Rev.* 1992; **60**: 291–319.
8. Shao J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* 1993; **88**: 486–494.
9. Hawkins D, Kraker J, Basak S, Mills D. QSPR checking and validation: a case study with hydroxy radical reaction rate constant. *SAR QSAR Env. Res.* 2008; **19**: 525–539.
10. Kraker J, Hawkins D, Basak S, Natarajan R, Mills D. Quantitative Structure-Activity Relationship (QSAR) modeling of juvenile hormone activity: comparison of validation procedures. *Chemom. Intell. Lab. Syst.* 2007; **87**: 33–42.
11. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J. Chemom.* 2009; **23**: 160–171.
12. Efron B. *The Jack-knife, the Bootstrap and Other Resampling Plans*. SIAM: Philadelphia, 1982.