The Validation of Toxicological Prediction Models

Graeme Archer,[1] Michael Balls,[1] Leon H. Bruner,[2] Rodger D. Curren,[3] Julia H. Fentem,[1] Hermann-Georg Holzhütter,[4] Manfred Liebsch,[5] David P. Lovell[6] and Jacqueline A. Southee[7]

[1]*ECVAM, JRC Environment Institute, 21020 Ispra (VA), Italy;* [2]*The Procter & Gamble Company, Health and Beauty Care Europe, Egham, Surrey TW20 9NW, UK;* [3]*Institute for In Vitro Sciences Inc., Suite 220, 21 Firstfield Road, Gaithersburg, MD 20878, USA;* [4]*Humboldt-Universität zu Berlin, Bereich Medizin (Charité), Institut für Biochemie, Mon Bijou Strasse 2a, 10117 Berlin, Germany;* [5]*ZEBET, Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin (BgVV), Diedersdorfer Weg 1, 12277 Berlin, Germany;* [6]*BIBRA International, Woodmansterne Road, Carshalton, Surrey SM5 4DS, UK;* [7]*Microbiological Associates Ltd, Stirling University Innovation Park, Stirling FK9 4NF, UK*

Summary — An alternative method is shown to consist of two parts: the test system itself; and a prediction model for converting *in vitro* endpoints into predictions of *in vivo* toxicity. For the alternative method to be relevant and reliable, it is important that its prediction model component is of high predictive power and is sufficiently robust against sources of data variability. In other words, the prediction model must be subjected to criticism, leading successful models to the state of confirmation. It is shown that there are certain circumstances in which a new prediction model may be introduced without the necessity to generate new test system data.

Key words: alternative method, model criticism, prediction model, validation.

Introduction

This is the report of a joint meeting of the European Centre for the Validation of Alternative Methods (ECVAM) task forces on biostatistics and prevalidation, which took place at ECVAM in February 1997. The aim of the meeting was to reach consensus concerning the development of prediction models within the validation process for alternatives to animal testing in toxicology.

The concept of validation has been greatly developed since the two Amden Reports were published (1, 2). A successful, and properly conducted, validation study is now acknowledged to be essential if an alternative method is to be accepted by the regulatory authorities. Together with work on the alternative test systems, some careful consideration of how we are to achieve the validation of alternative methods has been made (3).

A key advance has been the recognition that an *in vitro* test system must be reliable in two ways. Firstly, it must be possible to demonstrate that results obtained from testing individual substances in the same assay are reproducible across multiple laboratories and over time (2–3). Secondly, it must be possible to demonstrate that the predictions of toxicity from the alternative method are reproducible across appropriately defined sets of test substances (4). The device used to convert results from an alternative into a relevant prediction of toxicity has been called the prediction model. Recent work has demonstrated the importance to validation of having a well-defined and explicit prediction model prior to the start of the validation process. In fact, such models have always been implicit in the minds of toxicologists. For example, without the ability to convert a median neutral red uptake into a prediction of a Modified Maximum Average Draize Test Score (MMAS; for eye irritation), there would have been little point in first developing the neutral red assay.

An assay is considered relevant when its scientific significance and usefulness for a particular purpose have been established. The establishment of significance and usefulness is important because hazard predictions from scientifically credible alternative methods have a higher probability of being correct. In effect, the assessment of relevance answers the question, "Are the predictions obtained from a given alternative method 'good enough' for a defined purpose?".

It can therefore be seen that an alternative method (AM) has two main components: the *in vitro* test system (TS), and the prediction model (PM). This can be written symbolically as:

$$AM = PM \oplus TS$$

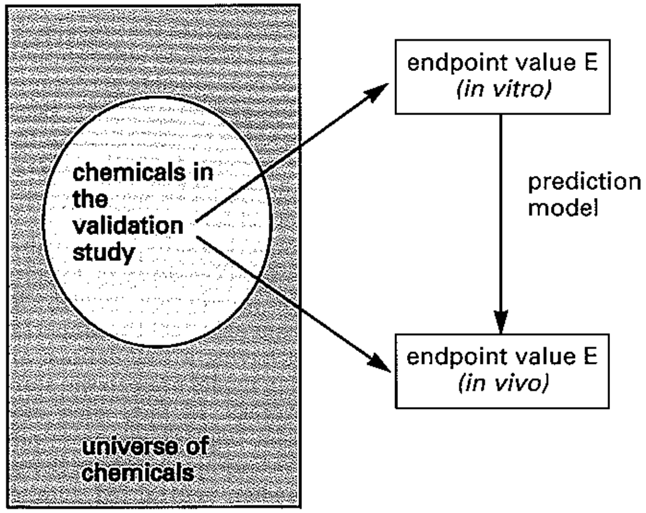
The main argument of this article is that the prediction model must be properly examined, as the test system is, during a validation study. This "model criticism" can lead to "model refinement" and, hopefully, to "model confirmation" (an approach advocated by Box *et al.* [5]). The following sections focus on why and how this can be achieved. Extensions to cases where alternative methods contain more than one *in vitro* test system, or make use of relevant structural information about the test substances, are also discussed.

In particular, the following questions are addressed.

1. Since a prediction model is a mathematical or statistical construct (rather than a real, physical object), how are its properties to be assessed? In particular, how can its predictive power be examined and tested?
2. Suppose a test system is evaluated in two separate experiments, A and B, and in both cases it is found to supply reproducible results. The only difference between the two experiments is in the set of test chemicals applied to the system. Supposing a prediction model has been evaluated and confirmed by using the results from experiment A, can we conclude that the model may also be judged suitable for use with chemicals of the type which were employed in experiment B?

Figure 1 demonstrates the arena in which these questions arise, by providing a schematic view of an alternative method. To identify the toxicity of a well-defined set of chemicals, one could choose either an *in vitro* or an *in vivo* method. For an alternative method to be useful, the endpoint from its test system (which must be reliably produced) must be converted by the prediction model into an estimate of the *in vivo* endpoint of interest. This report is concerned with how an accurate prediction model can best be established (question 1), and to determine the circumstances under which it could be used on chemicals which lie outside the set of chemicals upon which it has been confirmed (question 2).

Figure 1: A schematic representation of the role of the prediction model in an alternative method


chemicals in the validation study
universe of chemicals
endpoint value E (in vitro)
prediction model
endpoint value E (in vivo)

The Nature of the Prediction Model

In considering the nature of a prediction model, we build on the work of Bruner *et al.* (4), who list the following requirements for an adequate prediction model.

1. The specific purpose of the alternative method must be clearly specified.
2. A definition of all possible results from the alternative method must be supplied.
3. The prediction model must be able to specify how accurate its predictions are.
4. There must be a well-defined list of the classes of test substances for which the prediction model may be used. We return to this point in greater depth in the section *Information from Various Sources*.

A prediction model is likely to be a statistical or mathematical model, even if it is more commonly expressed as an algorithm, or rule. The term "model" is used in many circumstances and it is perhaps worthwhile to briefly dwell on the different types which exist. Statistical models are not physical laws (for example, Boyle's law linking pressure and temperature) because, while they may contain derived mechanistic relationships between variables, the actual parameterisaton of the model relies on the fitting of a postulated structure to experimental data, and cannot be deduced from a set of first principles. In particular, we are highly unlikely to be able to derive a causal relationship between *in vitro* and *in vivo* endpoints! Thus, while chemistry and biology may indicate relationships between *in vivo* and *in vitro* endpoints, and while statistical considerations of the nature of the endpoint data may have further structural implications, the *actual* model used for prediction must be derived from a conjunction of this structure with noisy experimental data, hence, prediction models will always have some sources of uncertainty. For example, an alternative method which uses a change in the electrical resistance of rat skin to predict corrosivity may have the following prediction model: "Identify any chemical for which rat skin electrical resistance falls beneath 5kOhms in a 24-hour period as corrosive". This initially appears to be non-statistical; however, why choose 5kOhms and not 6kOhms? This value is an example of a *parameter* within a particular model, which has been estimated in some way from some (possibly prevalidation) data. Another example would be the postulated relationship between median neutral red uptake values, which are a measure of cell viability, and the MMAS (4). When cell viability falls below a particular point, a certain level of eye irritation is inferred.

Note that we are *not* claiming that a prediction model is a mere statistical construct — the structure of the model ought to have a rational basis with respect to the biological phenomenon being modelled.

Whatever the prediction model, it should be noted that:

$$\text{model} = \text{structure} \oplus \text{parameters}.$$

In the electrical resistance example, the *structure* of the model is likely to be a logistic regression (or neural network), while the *parameter* is the cut-off value that separates *corrosive* from *non-corrosive*. The value "5" is an estimate of that parameter, and the true value will almost certainly never be known, even if the structure of the model is absolutely correct. Of course, it is unlikely that the structure of the model will ever be more than an approximation of the truth (that is, after all, the nature of modelling), so it is important that validation procedures address these sources of prediction uncertainty. In the neutral red uptake example, the *structure* is a linear relationship between NR50 values and MMAS, while the *parameters* are the estimates of the intercept and gradient of the line, which determine the exact correspondence between decline in cell viability and MMAS.

Attempts at developing non-statistical prediction models are likely to fail; if no probability distribution is postulated for a model, how can there be any confidence in the accuracy of the predictions from the model? For example, suppose a statistician selected a cut-off value of "26" in the electrical resistance example, because this value had worked well in experimental trials. How could he or she then estimate the variability of predictions from the model under repeated experimentation? This is one of the requirements for an adequate prediction model laid down by Bruner *et al.* (4). Without a statistical model, there is no way to sensibly answer such a question, nor to assess how robust the model is to mis-specification of these parameters. Even computer-intensive methods for inference (such as the bootstrap method, or a neural network) all postulate the existence of an identifiable model.

The difficulty with model building, as indicated above with the difference between a statistical model and a law of physics, is that it is a curious mixture of art and science. Statistical science can show the best fitting line (or hyperplane, or squiggly curve) with respect to any number of norms, once an underlying structural form has been chosen. It can indicate which of two competing models best explains a given set of data, and can do so without recourse to rigid linear relationships between the variables, or reliance on Gaussian distributional assumptions. However, it cannot tell us which of several competing structural forms to prefer, nor can it tell us which variables to consider for inclusion (although it can tell us if one is worth including after we have chosen it). Furthermore, there are an infinite number of models which will give equally good fits to any one set of data; how can we decide which one is the most correct, or most appropriate, for any given problem? The search for a best-fitting model is akin to the search for quality that marks the following passage (6):

> "To put it in more concrete terms: if you want to build a factory, or fix a motorcycle, or set a nation right without getting stuck, then classical, structured, subject–object knowledge, although necessary, isn't enough. You have to have some feeling for the quality of the work. You have to have a sense of what's good."

In other words, the prediction model needs to have a strong biological basis as well as being a good fit to data in the statistical sense.

Consider the following examples, both of which are statistical models.

1. If mean electrical resistance in three samples of rat skin drops beneath 5kOhms in a 2-hour period, the test chemical is corrosive.
2. If more than ten apples drop from the tree outside my office in a 60-minute period, the test chemical is corrosive.

We instinctively that the second model is invalid, and that an alternative method which employed it would not be relevant.

If the prediction model is to be at all useful, the model fitting (i.e. the process of choosing a structure and parameter estimates, using available data) must have been carried out in an appropriate manner. A cut-off value which was chosen merely because it "seemed effective" would not be appropriate, as we have seen. There are, of course, many ways to fit models (many are mentioned in the first report of the ECVAM biostatistics task force [7]), but the manner in which the fitting is carried out is not our focus here. However, the form of the model can have a bearing on the scope of validity of the prediction model.

The Validity of the Prediction Model

By a valid prediction model, we mean one that has sufficiently high predictive power for the *in vivo* endpoint of interest. The key point is that it should be scientifically acceptable as well as providing a good fit to the data, and that it should directly address the prediction of a well-defined biological response.

Bearing this in mind, it is possible to offer some examples of desirable properties for prediction models.

Primary observations rather than regulatory classifications

A prediction model can have the aim of predicting either a primary *in vivo* endpoint of toxicity (for example, quantifiable alterations of cells, tissues or organs observed in laboratory animals or in humans), or a classification of toxicity decided by a regulatory body. The electrical resistance example is an example of the latter.

Estimates of primary *in vivo* toxicity are more desirable than estimates of regulatory classifications, since the toxicity in a human exposed to the chemical is a real effect, arguably independent of regulatory assessment. In contrast, regulatory classifications are determined by political, or at best pseudo-scientific, considerations; they are at least one tier removed from the primary effect of interest (note that it is not suggested that the *in vivo* effects are free from variability). Moreover, many different regulatory classifications can be assessed from one set of *in vivo* predictions. For example, given four neutral red uptake 50% inhibition (IC50) concentrations for chemical X, what is the predicted probability that testing the same chemical in the Draize eye test will result in a corneal opacity score of, say, two?

A prediction model designed to answer this question is more relevant than one which makes a prediction of the probability that chemical X will fall into a class of toxicity defined by a regulatory authority (for example, that it should be classified as R36 according to EU risk phrase designations); in any case, this can be directly inferred from the prediction of corneal opacity.

Mirror the toxicological mechanism as closely as possible

The number of models which can be fitted to a set of data is vast, even if the obviously poorly fitting ones are discounted. To enhance the validity of the prediction model, it is advisable to use known chemical or biological information in choosing the model structure. For example, consider two continuous endpoints: the LD50 from an *in vivo* experiment, and the IC50 from an *in vitro* experiment. The aim of the prediction model is to predict LD50 values by using IC50 values. Standard practice would be to fit a simple linear regression to the values, and use the estimated parameters as the algorithm.

However, information is likely to be available on the structure of the chemicals to which the alternative method is to apply, for example, information regarding partition coefficients or acid dissociation constants. It is highly likely that incorporating such information into the prediction model will lead to more-effective predictions. Some work has been done on this in the quantitative structure-activity relationship (QSAR) literature (8), but it seems that most toxicologists view a prediction model and a QSAR as very distinct objects, when in fact both were designed with the same objective, i.e., to predict the activity of a chemical. A more integrated modelling strategy is needed; for example, a prediction model that utilises the results of an *in vitro* test as well as structure variables is more likely to be effective as it is most relevant, in the sense that it utilises more of the available information.

ECVAM is currently addressing such integrated testing strategies through another task force, while some work has been published by Blaauboer *et al.* (9) and Walum *et al.* (10).

The Precision of Predictions Obtained from a Prediction Model

We have discussed the structure and parameters of a prediction model. To discuss the precision of a prediction model, both components must be examined. Precision here means a standard of robustness in the predictions from the model; that the predictions are (sufficiently) accurate; and that well-defined estimates of error for any one prediction can be made. Checking a particular model via criticism of its components will lead, hopefully, to a confirmed prediction model.

Structure

The structure of the prediction model is the form of the relationship between the *in vitro* and *in vivo* endpoints. Is it a linear relationship? A quadratic one? A hideously non-linear one? Note that the true structural form can never be known, nor indeed can it always be assumed that one true model really exists. It is now "well-known" (11) that using the same set of data to derive a model's structure, and then to make inference about the population from which the data came using this model estimate, is likely to be misleading. It is therefore of benefit that the prediction model is defined before the validation study commences. However, a well-defined prediction model is emphatically not the same thing as the most precise prediction model. We return to this point below.

Parameters

The parameters are used to index the model within the family defined by its structure. For example, if a linear relationship links *in vitro* and *in vivo*, in the form:

in vivo = $\alpha + \beta$ *(in vitro)*

then the particular values chosen for α and β specify the particular form of the prediction model. Again, note that the true values of α and β will never be known and must be estimated on the basis of existing data. It is equally "well-known" that the more data available (assuming that all are generated from the same probability distribution), the more reliable the parameter estimates. Consider that a validation trial is likely to be one of the best-designed experiments of the test system and the prediction model. Thus, it is not beneficial that the prediction model is completely defined before the validation study commences, because new information in the data about the parameters is being wilfully ignored (the precision in the parameters will affect widths of confidence intervals and therefore judgements about toxicity).

Can data from a validation study be used to refine a prediction model once a study is completed? The answer is "yes", as long as this refinement occurs after the decision on the validity of the test system and the prediction model have been made, relative to the criteria established at the start of the study.

Post hoc prediction models: wise after the event?

There is considerable resistance in the toxicological community to the concept of *post hoc* fitting of prediction models by using data derived from the validation process itself. This stems from the belief that validation must be the last check before regulatory acceptance is sought for an alternative method and, as noted above, there are good statistical reasons for not basing model estimation and inference on the same set of data.

However, consider the concept of validity, in particular with respect to confirming the structure of a proposed prediction model. Is the model sufficiently tested by feeding in *in vitro* endpoints and comparing them with *in vivo* data? No, because a faulty prediction model could still make acceptable predictions, in some instances. Indeed, it is more likely that the prediction model considered in the validation process is only sub-optimal, rather than completely mis-specified. A slightly mis-specified structure (for example, linear rather than quadratic) is feasible if a sub-space of chemical classes is used during prevalidation (see the discussion of Figure 2 below). Therefore, it is recommended that the validation study data are used to check the defined prediction model by refitting a new model and comparing it with the prediction model. If the two are very similar (as would be expected), the structure of the prediction model has to that extent been *confirmed*. As Chatfield (12) said:

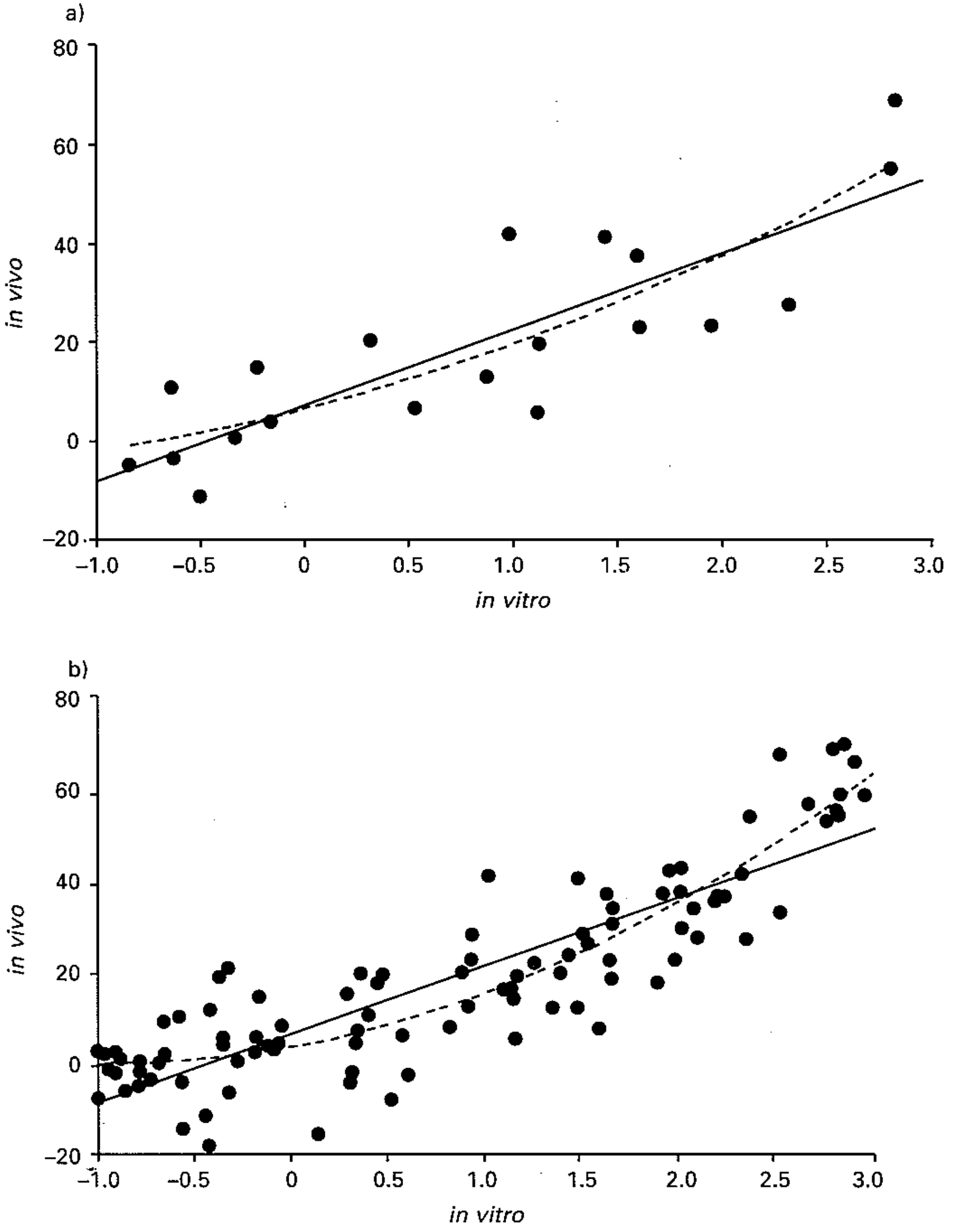
> "The only real validation of a statistical analysis, or of any statistical enquiry, is confirmation by independent observations (Anscombe [13], p. 6), and so model validation needs to be carried out on a *completely new* set of data."

Further, consider our comments on parametric uncertainty: if the chosen prediction model is considered to be structurally suitable, there still remains the problem of what to do about the information contained in the validation study data. Is it to be ignored? Since it can easily be shown that this would lead to more inaccurate predictions, surely this is not an option? Rather, it should be possible to *refine* a prediction model by updating the parameter estimates which index it within its (now confirmed) structure. Perhaps a suggestion for regulatory authorities can be made. As well as updating published protocols for animal experiments, *in vitro* techniques should also be subjected to periodic revision (there are a great many reasons why this should be the case, of which reducing parametric uncertainty in the prediction model is but one).

What should happen if the post-study prediction model is clearly superior?

It could happen that the post-study prediction model (PM_2) is not only structurally different to the existing prediction model (PM_1) but that, on the basis of the validation study data, it is also superior. What action should then be taken? One cannot simply replace PM_1 with PM_2 and declare the alternative method is validated. It has already been noted that it is unsafe to fit and assess a

Figure 2: Prevalidation and validation data sets with existing and *post hoc* prediction models



a) Prevalidation data set; b) validation data set.

Existing prediction model (PM_1, ——); post hoc prediction model (PM_2, - - -).

model with the same set of data. One way forward would be to confirm the superiority of PM_2 by applying it to the prevalidation data which should have been gathered before the full validation exercise commenced (14). This is *real* cross-validation; completely new, independent observations are used to confirm a postulated model structure. Alternatively, of course, classical cross-validation could be employed (15), by using various random subsets of the validation or prevalidation data sets to form new test sets for the new models.

Note that it is not the case that, because PM_1 was found to fit adequately, there is a "fault" with the prevalidation data. Consider Figure 2, which shows the case where the prevalidation data set suggests PM_1, a straight line fit, while with the addition of the validation study data set, a more appropriate structure — a curve — is suggested. The data were generated as follows: initially, the validation set was simulated to follow a true quadratic curve; then a subset of 20 chemicals were selected at random to be used as the prevalidation data set. On that set, both a straight line and a quadratic line were fitted. There is nothing wrong with the prevalidation data, but the important point is that PM_2, the curved line, provides an adequate fit to *both* sets, while PM_1, the straight line, is only able to accurately describe the restricted set. PM_2 is imbued with greater generality and therefore provides more-precise predictions of *in vivo* responses.

Of course, it is often the case that the prevalidation data set is the much larger of the two, in which case the labels on the diagrams in Figure 2 are reversed. However, the situation represented here can and does occur in the practice of validation studies, as some of the present authors can attest.

Conclusions concerning precision and validity

The post-study *refinement* of a published prediction model is perfectly acceptable, if it serves only to confirm the stated structure of the model and reduce uncertainty about parameter estimates, i.e., the refinement is carried out to increase the precision of the predictions from the model. If the structure of the post-study prediction model is very different, however, it is not advised that it merely replaces the extant one. Rather, very strong evidence against prediction model validity has been observed (even if the predictions of *in vivo* toxicity appear acceptable), and to this extent the prediction model being tested is a failure. In fact, whereas moving from the linear to the quadratic curve in Figure 2 is not such a dramatic shift, moving from a functional relationship such as that posited in Figure 2, to a highly nonlinear relationship requiring computer-intensive techniques of estimation, for example, could constitute a significant structural change. A possible way forward is to confirm the success of the new prediction model by applying it to existing prevalidation data, rather than to begin the entire validation process again from the beginning. Deciding whether a prediction model is worthy of refinement, as opposed to being completely structurally useless, should come under the aegis of the management team of a particular study. Such management teams are recommended to a priori draw up a table of minimum acceptable rates of correct classifications from the prediction models in their study, and to use this for comparison of the experimental results.

Information from Various Sources

The questions posed in the Introduction to this article can now be answered. Our discussion has repercussions for the case where the post-study prediction model is superior to that entered into the validation exercise, and for alternative methods which use more than one test system.

Consider the case where a test system has been shown to be reliable, giving highly reproducible results, in two separate "experiments" (experiments A and B; an experiment in this context means an exercise where a set of test chemicals are put through the test system in a manner concordant with the system's protocol).

Now suppose that either there was no prediction model in experiment B (perhaps the test system was still under development), or the prediction model which was used was found to be inadequate. In experiment A, not only was the test system good, but the prediction model (PM_1) was successful. Clearly, the alternative method defined by the conjunction of PM_1 and test system has been validated for chemicals which were tested in experiment A. However, in order to have an experimental method which covers the entire range of chemicals tested in both experiments, are further chemical tests needed to validate the conjunction of PM_1 with those classes of chemicals which were used in the assessment of the test system in experiment B?

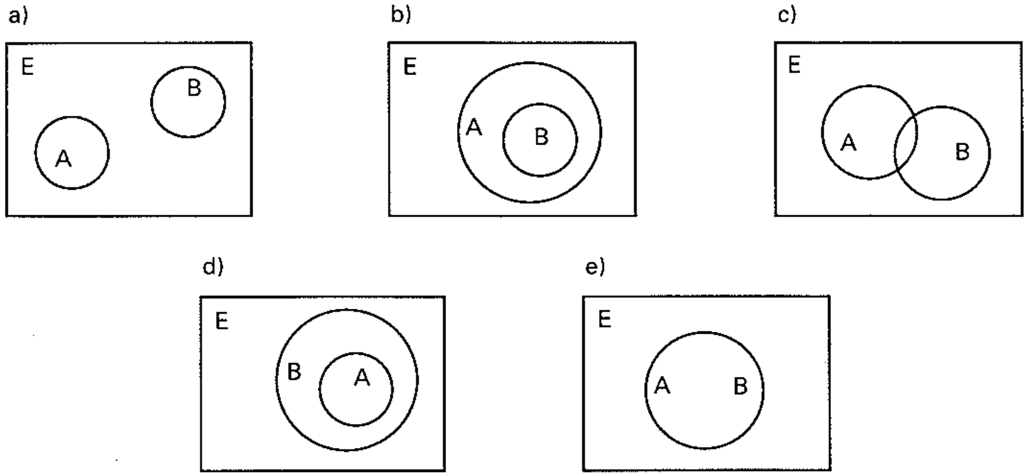
One of the requirements for a prediction model is that it specifies the classes of chemicals upon which it may validly be applied. Therefore, the alternative method has been validated for the classes of chemicals in experiment B which are common to those tested in experiment A.

We are assuming that the only possible difference between experiments A and B is in the choice of chemicals (they followed identical protocols). For example, experiment A may have tested only neutral organic and surfactant chemicals, while experiment B may have tested neutral organics and phenols. There are, therefore, five possibilities to consider for the relationship between A and B, shown as Venn diagrams in Figure 3.

In Figure 3a, there is no overlapping of the chemical classes tested in the two experiments, while Figure 3b represents the case where experiment B tests are a subset of the classes of experiment A. In Figure 3c, there is an intersection of chemicals tested, but still some different testing in both experiments. Figure 3d is the reverse of Figure 3b; the classes in experiment A were a subset of those in experiment B. Finally, in Figure 3e, the classes tested in both experiments were identical. Note that identical classes do not imply identical test substances (as an aside, it can be envisaged that QSAR methodology could be brought to bear in determining which chemicals belonged to different classes).

To return to the equation, $AM = PM_1 \oplus TS$, consideration of these diagrams demonstrates the extent to which the alternative method has been validated, that is, the range of chemical classes over which the alternative method can be said to be reliable and relevant. If the situation is as shown in Fig-

Figure 3: Venn diagrams of the five possible relationships between chemicals tested in hypothetical experiments A and B



a) Chemical classes in experiments A and B are mutually exclusive; b) chemical classes from experiment B are a subset of chemical classes from experiment A; c) chemical classes from experiments A and B intersect; d) chemical classes from experiment A are a subset of chemical classes from experiment B; and e) chemical classes in experiments A and B are identical.

A = chemical classes from experiment A; B = chemical classes from experiment B; and E = all chemicals.

ure 3e, the alternative method has been as well-validated as if a special study had been carried out for testing PM_1 and the chemicals which were used in experiment B in unison, for the prediction model was tested on an identical class of chemicals as has been the test system. Conversely, if experiments A and B took the form of Figure 3a, the alternative method, defined by $PM_1 \oplus TS$, has not been validated for the classes of chemicals in experiment B, as there is no intersection between the classes of chemicals in the two experiments.

In general, if a prediction model is validated in an experiment, independently of another involving the same test system, the alternative method defined by the juncture of the prediction model and the test system can be said to have been validated where the chemical classes from the two experiments intersect. This can be termed the "scope of validity" for the alternative method. This is perhaps a complicated manner in which to state the obvious!

Now let us consider the following post-study scenario. Suppose that there are two test systems, TS_X and TS_Y, which are (perhaps radically) different in their scientific basis, but which are both designed to be coupled with prediction models that have a common *in vivo* endpoint. Suppose that both TS_X and TS_Y were found to be reliable in the validation study, but their associated prediction models have not been very successful in predicting *in vivo* toxicity. However, a post-study prediction model, PM_{XY}, has been found, which utilises endpoints from both TS_X and TS_Y and which is successful when applied to the validation data. This is encouraging, but once again the problem of over-fitted models is raised: PM_{XY} cannot be defined and evaluated on the same (validation) data set. Recourse can be made to the prevalidation data. There should be a good stock of this for TS_X and TS_Y, which can be used to confirm post-study PM_{XY}. The discussion on chemical class intersection completes the validation procedure. The new alternative method, AM_{XY}, is defined by:

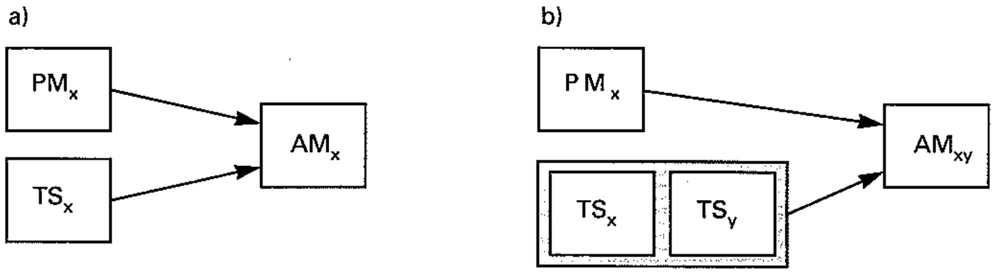
$$AM_{XY} = PM_{XY} \oplus TS_X \oplus TS_Y.$$

This is validated for chemicals belonging to the *intersection* of the chemical classes used in the prevalidation and validation of TS_X and TS_Y.

In Figure 4, both the original alternative method for only TS_X and PM_X, and the new AM_{XY}, are represented. Figure 4a is an example of the most straightforward alternative method, built from one test system and one prediction model. Figure 4b is a representation of the more complicated AM_{XY}. The Venn diagram in Figure 5 represents the scope of validity for AM_{XY}, that is, the shaded intersection of the two sets of test chemicals.

Such diagrams are useful devices for demonstrating the structure of any particular alternative method. A validation trial could choose to investigate one or other arm of the alternative method (or both at once), as long as it is made clear, at the very beginning, to which arm the results will apply.

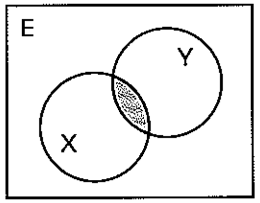
Figure 4: A schematic representation of alternative methods



a) Alternative method X (AM_X), consisting of test system X (TS_X) and prediction method X (PM_X).

b) New alternative method XY (AM_{XY}), consisting of test systems X and Y (TS_X, TS_Y) and a new prediction model (PM_{XY}).

Figure 5: Venn diagram of the scope of validity for alternative method XY (AM_{XY})



Shaded area shows scope of validity for AM_{XY}. AM_{XY} is schematically shown in Figure 4b.

X = chemical class X; Y = chemical class Y; and E = all chemicals.

Conclusion

It is as important to validate the prediction model as it is to establish the reliability of a test system, for a precise and robust prediction model is required to give relevance to the alternative method, defined as the conjunction of the test system and the prediction model. If the performance of the alternative method meets the criteria defined prior to the start of the study, it should be considered valid. If further consideration of new data shows that small adjustments in the parameters of a valid model will improve the precision of predictions, post-study refinement is acceptable.

If the alternative method fails to meet the criteria, the method must be declared not valid. If this occurs, but results from the study indicate that modifications (to the structure) of the prediction model could improve the predictive capacity, these changes must be defined. The alternative method can then be retested in a subsequent validation study using an independent set of test chemicals. Whether the data come from the prevalidation test sets or anywhere else is irrelevant as long as they are independent of the data set used to develop this new prediction model; thus, further laboratory work involving the test system might not be required.

References

1. Balls, M., Blaauboer, B.J., Brusick, D., Frazier, J., Lamb, D., Pemberton, M., Reinhardt, C., Roberfroid, M., Rosenkranz, H., Schmid, B., Spielmann, H., Stammati, A-L. & Walum, E. (1990). Report and recommendations of the CAAT/ERGATT workshop on the validation of toxicity test procedures. *ATLA* **18**, 303–337.
2. Balls, M., Blaauboer, B., Fentem, J.H., Bruner, L., Combes, R.D., Ekwall, B., Fielder, R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski, D., Spielmann, H. & Zucco, F. (1995). Practical aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM workshop 5. *ATLA* **23**, 129–147.
3. Balls, M. & Fentem, J.H. (1997). Progress toward the validation of alternative tests. *ATLA* **25**, 33–43.
4. Bruner, L.H., Carr, G.J., Chamberlain, M. & Curren, R.D. (1996). Validation of alternative methods for toxicity testing. *Toxicology in Vitro* **10**, 479–501.
5. Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters,* 653 pp. New York: John Wiley.
6. Pirsig, R.M. (1974). *Zen and the Art of Motorcycle Maintenance*, 424 pp. London: Vintage.
7. Holzhütter, H-G., Archer, G., Dami, N., Lovell, D.P., Saltelli, A. & Sjöström, M. (1996). Recommendations for the application of biostatistical methods during the development and validation of alternative toxicological methods. ECVAM biostatistics task force report 1. *ATLA* **24**, 511–530.
8. Barratt, M.D., Dixit, M.B. & Jones, P.A. (1996). The use of *in vitro* cytotoxicity measurements in QSAR methods for the prediction of the skin corrosivity potential of acids. *Toxicology in Vitro* **10**, 283–290.
9. Blaauboer, B.J., Balls, M., Bianchi, V., Bolcsfoldi, G., Guillouzo, A., Moore, G.A., Odland, L., Reinhardt, C.A., Spielmann, H. & Walum, E. (1994). The ECITTS integrated toxicity testing scheme: the application of *in vitro* test systems to the hazard assessment of chemicals. *Toxicology in Vitro* **8**, 845–846.
10. Walum, E., Balls, M., Bianchi, B., Blaauboer, B.J., Bolcsfoldi, G., Guillouzo, A., Moore, G.A., Odland, L., Reinhardt, C. & Spielmann, H. (1992). ECITTS: an integrated approach to the application of *in vitro* test systems to the hazard assessment of chemicals. *ATLA* **20**, 406–428.
11. Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika* **79**, 741–746.
12. Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A* **158**, 419–466.

13. Anscombe, F.J. (1967). Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *Journal of the Royal Statistical Society, Series B* **29**, 1–52.
14. Curren, R.C., Southee, J.A., Spielmann, H., Liebsch, M., Fentem, J.H. & Balls, M. (1995). The role of prevalidation in the development, validation and acceptance of alternative methods. *ATLA* **23**, 211–217.
15. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36**, 111–147.