Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs

Lennart Eriksson,[1] Joanna Jaworska,[2] Andrew P. Worth,[3] Mark T.D. Cronin,[4] Robert M. McDowell,[5] and Paola Gramatica[6]

[1]Umetrics, Umeå, Sweden; [2]Procter & Gamble Eurocor, Central Product Safety, Strombeek-Bever, Belgium; [3]European Chemicals Bureau, Institute for Health & Consumer Protection, Joint Research Centre, European Commission, Ispra, Italy; [4]School of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool, United Kingdom; [5]U.S. Department of Agriculture, Animal and Plant Health Inspection Service, Risk Analysis Systems, Riverdale, Maryland, USA; [6]QSAR and Environmental Chemistry Research Unit, Department of Structural and Functional Biology, Insubria University, Varese, Italy

This article provides an overview of methods for reliability assessment of quantitative structure–activity relationship (QSAR) models in the context of regulatory acceptance of human health and environmental QSARs. Useful diagnostic tools and data analytical approaches are highlighted and exemplified. Particular emphasis is given to the question of how to define the applicability borders of a QSAR and how to estimate parameter and prediction uncertainty. The article ends with a discussion regarding QSAR acceptability criteria. This discussion contains a list of recommended acceptability criteria, and we give reference values for important QSAR performance statistics. Finally, we emphasize that rigorous and independent validation of QSARs is an essential step toward their regulatory acceptance and implementation. *Key words:* QSAR acceptability criteria, QSAR applicability domain, QSAR reliability, QSAR uncertainty estimation, QSAR validation. *Environ Health Perspect* 111:1361–1375 (2003). doi:10.1289/ehp.5758 available via *http://dx.doi.org/* [Online 5 February 2003]

Introduction

General Considerations

Quantitative structure–activity relationships (QSARs) are mathematical models approximating the often complex relationships between chemical properties and biological activities of compounds. Common objectives of such models are *a*) to allow prediction of biological activity of untested and sometimes yet unavailable compounds and *b*) to extract clues of which chemical properties of compounds are likely determinants for their biological activities. It is convenient to distinguish between QSARs and SARs: QSARs are typically quantitative in nature, producing categorical or continuous prediction scales; SARs are qualitative in nature, often occurring in the form of structural alerts that include molecular substructures or fragment counts related to the presence or absence of biological activity.

In this article, we review QSARs. The most common techniques for establishing QSARs are based on regression analysis, neural nets, and classification approaches. Among the regression-based approaches, the methods of multiple linear regression (MLR) and partial least squares (PLS) regression are prime examples. Examples of classification methods involve, for example, discriminant analysis and decision trees. It is important to observe that classification is a central concept also in regression-based QSARs. A molecule that is not satisfactorily classified in a model—that is, it does not "fit" the model—should be handled with care, and the model's predictions should be considered with some skepticism. Hence, methods and tools for classification are ubiquitous in QSARs, regardless of the final form of the "equation."

QSARs are increasingly used by authorities, industries, and other institutions for assessing the risks of chemicals released to the environment (Anonymous 1995, 1999). An important reason for this is the increasing awareness that completing even the most basic biological testing of compounds of concern would take decades. Therefore, predictive models (PMs) such as QSARs are necessary for aiding in chemicals management because they may considerably reduce costs, avoid animal testing, and speed up managerial decisions. In addition, safety of new chemicals, often already in the preproduction phase, can be assessed via QSARs (Anonymous 1995, 1999; Wahlström 1988). This may guide the design of compounds with fewer unwanted side effects by optimizing their relevant properties. However, these potential benefits of QSARs can be fulfilled only if QSAR results are accepted at the regulatory level. The decision to accept QSAR results relies on assessing the reliability and uncertainty of the predictions as well as assessing the applicable domain of a QSAR.

Analogy Models

The goal in any QSAR modeling is to obtain the mathematical expression that best portrays the relationship between chemistry and biology. To adequately describe the often complex nature of such phenomena, it is often necessary to use a battery of relevant and consistent chemical descriptors (Dunn 1989; Eriksson et al. 2001; Wold and Dunn 1983). The assumption, or expectation, is then that the factors governing the events in a biological test system are represented in the multitude of descriptors characterizing the compounds. In other words, within a series of compounds—in which biological activity is expressed via the same mechanism—it is anticipated that a small change in chemical structure will be accompanied by a proportionally small shift in biological activity, and that the set of descriptors will reveal these analogies. Hence, QSARs are sometimes referred to as analogy models (Eriksson et al. 2001; Wold and Dunn 1983).

Analogy models can be regarded as linearizations of the real, complicated SARs. Wold and Dunn (1983) have shown that such analogy models normally have local validity only, that is, can embrace only compounds with similar chemical and biological data. It is noted, however, that the substances must be disparate enough to cause some systematic change in biological activity.

The nature of the biological response variable under study has a strong impact on the degree of chemical diversity that can be accommodated by a QSAR model; that is, there is a trade-off between chemical diversity in the training set and complexity of the biological response variable (Wold and Dunn 1983; Eriksson et al. 2001). For an endpoint variable where measured data involve a specific and selective mechanism, it is expected that the resulting QSAR model cannot tolerate too much structural diversity in chemicals (Anonymous 1995, 1999). On the other hand, in less complicated cases, dealing with less "demanding" biological response variables, for example, acute toxicity of narcotics to aquatic organisms, QSAR models are usually possible for a much broader and more diverse set of chemical structures.

The Role of Pattern Recognition in QSARs

Pattern recognition (PARC) is often described as a procedure for formulating rules of classification (Albano et al. 1978; Wold et al. 1983) in multivariate data. PARC has been used in a wide variety of applications such as

This article is part of the mini-monograph "Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints."

Address correspondence to L. Eriksson, Umetrics AB, POB 7960, 907 19 Umeå, Sweden. Telephone: 46-90-184852. Fax: 46-90-184899. E-mail: lennart.eriksson@umetrics.com

The authors declare they have no conflict of interest.

Received 2 May 2002; accepted 3 February 2003.

analytical chemistry, food research, and process monitoring in manufacturing. PARC methods are useful also in QSARs (Wold and Dunn 1983). Based on a set of given classes, each of which contains a number of observations (in QSARs, compounds) mapped by a multitude of variables, guidelines, and rules are developed that make it possible to classify new observations (compounds) as similar or dissimilar to the members of the existing classes.

Experience shows that nature often seems to organize itself in a clustered, rather discontinuous way. Inside a class or a cluster, the observations (compounds) are rather similar to each other, so if we know the class membership of an observation, we can potentially infer a great deal about it. The similarity among observations within each class is considerably greater than among observations of different classes. This is the basis for the principle of analogy. If we know that a compound is a hydrocarbon, for instance, we can confidently predict how the compound reacts or fails to react with various “reagents” because we know from experience that almost all hydrocarbons behave similarly, analogously, when subjected to various “treatments.”

It is therefore often practical to formulate a QSAR problem in terms of similarities and classes. One tries to find a battery of easily accessible properties (variables) that can be used to predict the class of an unknown observation (compound). One then infers that all observations within a class behave similarly and that there are no outliers or further subclustering endangering the foundation of the class model. Once such information is known, it is also possible to determine which among existing—perhaps competing—QSAR models will best accommodate a candidate chemical for which prediction of biological and environmental data is sought.

Scope of Review

The objective of this article is to review existing methods for assessing the reliability and uncertainty of QSARs, particularly regarding predictive power and applicability domain. In so doing, the objective is also to distill some indicators that can be used as acceptability criteria. In the section, “Conditions for Applicability and Validity of QSARs,” we outline basic conditions for the applicability of QSARs. In ”Modeling Techniques” we review common modeling techniques in QSARs, with emphasis on regression-based methods. In “Assessing/Enhancing Model Reliability, Interpretability, and Predictive Power,” we describe various tools that aid the development and use of QSARs. In “Bayesian Methods for Reliability Testing,” we consider Bayesian approaches and their applicability in QSAR reliability assessment, and, last, in the “Discussion,” we provide concluding remarks with recommendations for acceptability criteria.

We make very clear that we are addressing important matters of QSARs from a statistical perspective. Thus, the main focus lies on discussing methods, procedures, and diagnostic tools—mostly statistical in nature—aiding us in developing statistically and informationally sound QSARs. However, this very strong emphasis on data analytical aspects of QSARs does not mean that we refrain entirely from touching upon related and important items that deal with, for example, compilation of chemical and biological data, configuration of data tables, and so forth. It should be emphasized, however, that we do not intend to delve deep into detailed and practical issues regarding procedures for gathering the necessary chemical, biological, and toxicologic data.

Conditions for Applicability and Validity of QSARs

Homogeneity

Any data analysis, including QSAR modeling, is based on an assumption of homogeneity and absence of influential outliers (Wold et al. 1993; Eriksson and Johansson 1996). This means that the investigated system, that is, series of compounds, must be in a similar “state”—have rather similar properties—throughout the investigation, and the mechanism of influence of *X* on *Y* must be the same. This, in turn, corresponds to having some limits on the variability and diversity of *X* and *Y*. These limits may be wide if the biological activity is unspecific (e.g., acute toxicity to fish for narcotic chemicals), or narrow if the biological endpoint involves a very specific mechanism of action (e.g., binding of substrates to the active site of an enzyme).

Hence, it is essential that the data analysis provide diagnostics about how well these assumptions indeed are fulfilled. Much of the recent progress in applied statistics has concerned diagnostics, and many of these diagnostics can be used also in QSAR modeling as discussed later.

In many cases, QSAR modeling in risk assessment involves large databases of clustered compounds. Here the term “clustered” corresponds to a data set in which several classes of chemical compounds are encountered. These classes may be partially overlapped, barely separated, or completely resolved in the chemical descriptor (X-) space and/or biological property (Y-) space of the compounds in question. To conduct proper QSAR modeling, it is important to understand the nature of the clustering that occurs.

The extent to which data are clustered will be a function of the compounds and descriptors chosen, and can be checked by simply plotting the data and/or model parameters. In the ideal case, compounds will have an even spread in such plots. Moreover, there should be no influential outliers or strong clustering. If there is strong clustering in the data, it is often not realistic to fit only one model. Such a model would be able to describe only systematic variation among the groups and would be unable to resolve what is happening within a group. We also note that, from a modeling point of view, too, severe clustering will violate the assumption of homogeneity; that is, if a data set is clustered with large separation between groups, it no longer has a homogeneous distribution.

Therefore, with selective and specific biological or environmental responses, and a strongly clustered data set—a chemical property space containing several dense regions (clusters) of compounds with empty space between—it is often appropriate to treat each cluster/class independently and make separate QSAR models for each homogeneous cluster (Andersson et al. 2000; Eriksson et al. 2000a).

However, with nonspecific responses, often resulting from measurement in aquatic environments and with less strong clustering in the chemical properties, that is, clusters that are partially overlapped or barely resolved from one another in chemical space, the approach is a bit more complicated. Although a single QSAR model is still conceivable, care must be exercised to assure all chemical classes are represented in the training set (Andersson et al. 2000; Eriksson et al. 2000a). Otherwise, there is an apparent risk that small clusters with few members will not be represented in the final training set.

Representativity

As should be apparent from the discussion above, the selection of the training set is crucially important in QSAR analysis. A representative selection of compounds that well span the chemical domain of interest should be included in this set. One way to accomplish a representative training set is through multivariate design (Wold et al. 1986). This methodology is also frequently used in medicinal chemistry and combinatorial approaches and is known as statistical molecular design (SMD). It results in a test series of compounds in which all major structural and chemical properties are systematically varied at the same time (Giraud et al. 2000; Linusson et al. 2000).

A point of some controversy is how to define the chemical space appropriately. This is not a trivial issue. Because it is often difficult to know beforehand exactly which type and combination of chemical descriptors will be found useful in the QSAR modeling, the general advice given is to include a broad and stable set of descriptors.

The ensuing data analysis will then reveal whether the data set contains groups, outliers, and so forth, and care must then be exercised to modify the data set accordingly. Moreover, QSAR practitioners are sometimes anxious regarding the consequences of forgetting to include important chemical descriptors when compiling the initial set of descriptors. Frequently, however, this is not a big problem. If extra variables are added to the data set during the QSAR analysis, and if these are few compared with the total number of descriptors used, the structure of the training set in terms of its latent variables usually is little affected.

Moreover, it is important to understand the range of validity of the QSAR model-to-be, both in terms of the range of biological response data within which it will predict reliably, and also in terms of the type of chemical structure on which it is based. Diagnostic tools aiding us in the assessment of such model validity ranges are discussed.

Demands on the *X*-Data (Chemical Descriptors) and *Y*-Data (Biological Responses)

The intuitive belief of many environmental chemists and toxicologists is that measuring many variables provides more information about the chemical and biological properties of compounds than measuring just a few variables. Indeed, a rich description of chemical properties of compounds will facilitate the detection of groups (classes) of compounds with markedly different properties and help in unraveling chemical outliers. Outliers are compounds that do not fit a QSAR. It is important not to simply mechanically delete such compounds from a data set; rather, they should be analyzed carefully because their existence might lead to new, unexpected discoveries.

The compilation of data for use in QSARs requires consideration of some important aspects. First of all, because all our QSAR modeling efforts rest critically on the assumption of chemical similarity and biological homogeneity of compounds, we must analyze data that are rich enough to allow an adequate testing of this important assumption. This means that we must use chemical descriptors that are meaningful, interpretable, and reversible.

Descriptors that are often found useful in QSARs mirror fundamental physicochemical factors that in some way relate to the biological endpoint(s) under study. Examples of such molecular properties are hydrophobicity, steric and electronic properties, molecular weight, pK_a, and so forth. These descriptors provide valuable insight into plausible mechanistic properties. It is also desirable that the chemical description be reversible. It must be possible to convert model information into understandable chemical properties. For a deeper treatment of chemical descriptors and their use in QSARs, the reader is advised to consult the literature (e.g., Andersson et al. 2000; Cronin and Schultz 2003).

Furthermore, as emphasized by Cronin and Schultz (2003), knowledge about the biological data is essential in QSARs:

> Reliable data are required to build reliable predictive models. In terms of biological activities, such data should ideally be measured by a single protocol, ideally even the same laboratory and by the same workers. High quality biological data will have lower experimental error associated with them. Biological data should ideally be from well standardized assays, with a clear and unambiguous endpoint.

The article of Cronin and Schultz (2003) also discusses in depth the importance of appreciating the quality of biological data and of knowing the uncertainty with which the biological data were measured.

Interestingly, QSAR analysis may involve modeling of more than one endpoint, that is, a matrix (*Y*) of several end points. This will lead to the determination of biological response profiles of compounds (Nendza and Müller 2000). Measurement of multivariate biological data leads to statistically beneficial properties of the QSAR and improved possibilities of exploring the biological similarity of the studied substances. The absence of outliers in multivariate biological data is a very valuable indication of homogeneity of the biological response profiles among the compounds (Eriksson et al. 2001, 2002).

The use of multiple endpoints is becoming increasingly widespread in QSARs, in both drug design and environmental sciences (Deneer et al. 1987, 1989; Nendza and Müller 2000; Sjöström et al. 1997; Verhaar et al. 1994). And, as discussed above, a multitude of chemical descriptors is often favorable and tends to stabilize the description of the chemical properties of the compounds.

Modeling Techniques

Multiple Linear Regression

MLR is the classical approach to regression problems in QSARs. MLR assumes the predictor variables, normally called *X*, to be mathematically independent (orthogonal). Mathematical independence means that the rank of *X* is *K* (the number of *X*-variables).

A limitation of MLR is the sensitivity to correlated descriptors. One practical workaround is to use long and lean data matrices—matrices where the number of compounds substantially exceeds the number of chemical descriptors—where interrelatedness among variables usually drops. It has been recommended that the ratio of compounds to variables should be at least 5 (Topliss and Edwards 1979). We note that one way to introduce orthogonality or near-orthogonality among the *X*-variables is through SMD.

MLR is satisfactorily applied in QSAR studies if the main problem of the selection of variables is faced and solved.

MLR is usually used to fit the regression model (Equation 1), which models a response variable, *y*, as a linear combination of the *X*-variables, with the coefficients *b*. The deviations between the data (*y*) and the model (*Xb*) are called residuals, and are denoted by *e*:

$$y = Xb + e \qquad [1]$$

For many response variables (columns in the response matrix *Y*), regression normally forms one model for each of the *M y*-variables, that is, *M* separate models.

If MLR is applied to data sets exhibiting collinearities among the *X*-variables, the calculated regression coefficients get unstable and their interpretability breaks down (Draper and Smith 1981; Lindgren 1994; Topliss and Edwards 1979). For example, certain coefficients may be much larger than expected, or they may even have the wrong sign (Eriksson et al. 1995; Lindgren 1994; Mullet 1976).

Another key feature of MLR is that it exhausts the *X*-matrix, that is, uses all (100%) of its variance (i.e., there will be no *X*-matrix error term in the regression model). Hence, it is assumed that the *X*-variables are exact and completely (100%) relevant for the modeling of *Y*.

Other Approaches

Multivariate projection methods such as principal component analysis (PCA), principal component regression (PCR), and PLS are other approaches that are increasingly used in QSAR analysis in the environmental sciences (Langer 1994; Sjöström et al. 1997; Tosato et al. 1992; Tysklind et al. 1995; Verhaar et al. 1994). These methods are particularly apt when the number of variables equals or exceeds the number of compounds. This is because projections to latent variables in multivariate space tend to become more distinct and stable as more variables are involved (Eriksson et al. 2001; Höskuldsson 1996; Lindgren 1994; Wold et al. 1993).

Geometrically, PCA, PCR, PLS, and similar methods can be seen as the projection of the observation points (compounds) in variable space down on an A-dimensional hyperplane. The positions of the observation points on this hyperplane are given by the scores, and the orientation of the plane in relation to the original variables is indicated by the loadings. In contrast to MLR, PLS and similar approaches do not exhaust the *X*-matrix; that is, they do not assume that the *X*-variables are exact and 100% relevant for modeling of *Y*.

Some other methods are canonical correspondence analysis (CCA), correspondence analysis scaling (for discrete data), redundancy analysis, and ridge regression (Jackson 1991; Jongman et al. 1987).

MLR, PLS, and the other methods discussed above are usually applied to data sets where a linear relationship between *X* and *Y* is anticipated. However, there are also many other methods that are used in the analysis of nonlinear QSAR data, for example, neural networks (Burden et al. 1997), nonlinear versions of genetic algorithms (Vankeerberghen et al. 1995), and nonlinear extensions of PLS (Eriksson et al. 2000a; Martin et al. 1995; Wold 1992). All these methods contain more adjustable model parameters than do linear modeling techniques. As a consequence, nonlinear modeling methods are usually very flexible and adapt to almost anything, including outliers, inhomogeneities, discontinuities, and other anomalies in the data. Because of the high degree of flexibility of such methods, very many observations (compounds) are required for these techniques to work reliably and produce stable models.

A recent article by Worth and Cronin (2003) describes the use of alternative techniques in QSARs, such as discriminant analysis, logistic regression, and classification tree analysis. The reader is referred to this article for a more in-depth discussion of the classification problem and how to categorize compounds as active/inactive or potent/nonpotent using these approaches.

Assessing/Enhancing Model Reliability, Interpretability, and Predictive Power

A QSAR analyst must master many elements of data analysis. There are many tools and diagnostics available that will give better, more reliable, and more useful PMs. In this section we provide an overview of some of these tools and diagnostics.

Preprocessing Techniques

Scaling and centering. Pretreatment of measured data is carried out to reshape ("transform") the data to facilitate data analysis and model interpretation. The two most common preprocessing procedures are centering and scaling (Eriksson et al. 2001). Subtracting the mean (mean-centering) facilitates model interpretation and may in certain situations also remove some numerical instability.

An initial scaling of data, often to a variance of 1 for each variable (unit-variance scaling), is done to ensure that all variables have the same chance to influence a regression model (Eriksson et al. 2001). This type of preprocessing is especially useful when the variables considered are of different origin and display considerably different numerical range. Without any scaling, variables with large numerical range would otherwise dominate over variables with small numerical range. Figure 1 shows an example involving two variables where one variable can be made to dominate over the other when scaling is not done appropriately.

One additional approach of considerable interest for the future is called Pareto scaling (Eriksson et al. 2001), whereby each variable is given a variance equal to its standard deviation rather than unit variance. It can be seen as a compromise between no scaling (risk: "small" variables will be masked by "large" variables) and unit-variance scaling (risk: noise is inflated because noisy variables are up-weighted).

In summary, scaling can be done in many different ways, depending on the modeling objectives and the level of prior knowledge about the properties of the data. Also, if the uncertainties of the *X*- and *Y*-data are estimable, such information may be used to modify the scaling weights. For instance, if in a given situation it is known that the standard deviation of an *X*-variable is three times higher than that for any other *X*-variable, a down-weighting by one-third would seem reasonable, thus giving this *X*-variable a variance of one-third rather than unity.

Data correction and compression. Data pretreatment often has wider scope than just scaling and centering. In spectroscopically based QSAR applications, occurring mainly in pharmaceutical industry, spectral data are often transformed to remove undesired systematic behavior ("signal correction") (Wold and Josefson 2000). Such undesired variation may arise from light-scattering effects, baseline drift, nonlinearities, and so forth, which influence the shape of the spectral data without really being relevant to the *Y*-data modeled. Therefore, there is an interest of "correcting" the spectral data and removing from the *X*-matrix the variation that does not relate to the *Y*-data. The "corrected" or "filtered" *X*-matrix then contains the variation that correlate with *Y*, and hence the QSAR model is better focused.

Signal correction improves the interpretability of a QSAR model and may also improve its predictive power. A facilitated transfer of a model from one site to another, so-called calibration transfer, may also be the result (Sjöblom et al. 1998). Furthermore, when very large sets of spectral data are investigated, the pretreatment phase may also involve measures to reduce the size of the data material (signal compression), for instance, by using wavelet compression (Wold and Josefson 2000). For a discussion of useful correction and compression approaches, see Eriksson et al. (2001).

Transformations. Another situation for preprocessing of raw data is when a variable contains one or a few extreme measurements that may unduly influence model building. Consider Figure 2A, which shows the histogram of a variable, Var1. One out of the 40 measurements in this variable is substantially larger than the others. If this extreme measurement is not manipulated in some way

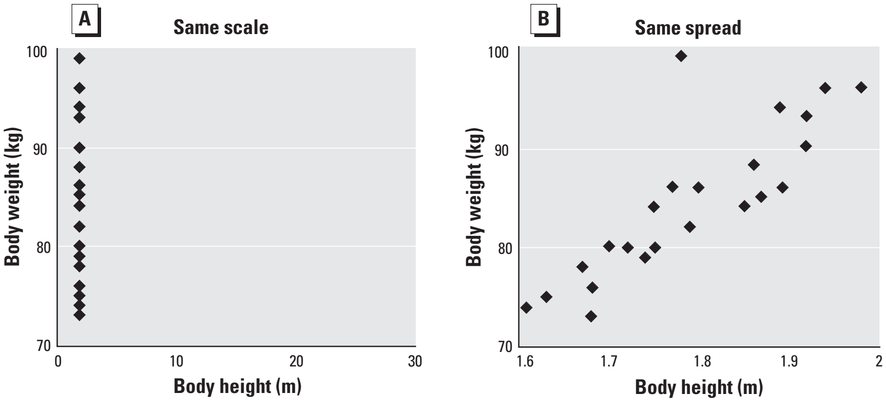

Figure 1. (*A*) Scatterplot of body weight versus body height of 23 individuals (22 football players and one referee). The data pattern is dominated by the influence of body weight. The two variables have been given the same scale. (*B*) Scatterplot of body weight against body height of 23 individuals. Now, the variables are given equal importance by displaying them according to the same spread. An outlier, a deviating individual, the referee of the game, is now discernible.

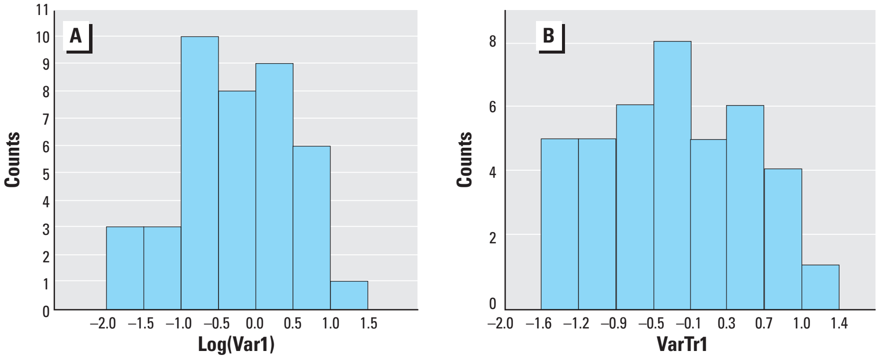

Figure 2. (*A*) Histogram of a nontransformed variable Var1. (*B*) Histogram of variable Var1 after log transformation.

before the data analysis, it will exert a large influence (have high leverage) on the model and dominate over the other measurements. A simple logarithmic transformation will in this case remedy the situation (Figure 2B). If the transformation does not increase the model's goodness of prediction, it should be avoided.

Informative Model Parameters

Depending on the data analytical technique used, QSAR analysis will result in a set of model parameters that is useful in the interpretation phase. With straightforward MLR, a regression equation consisting of coefficients is produced. These coefficients have an intuitively simple and therefore appealing meaning. But, one should be aware that—depending on the choice of regression method—there are other model parameters and diagnostics available that also deserve attention when interpreting a QSAR model. Our goal in this subsection is to highlight a few of these parameters and diagnostics. In so doing, we will use two data sets drawn from the literature.

Interpretation with emphasis placed on regression coefficients,* Y*-residuals, and model performance statistics. The first data set deals with toxicity data (ICG_{50}, concentration causing 50% growth inhibition to *Tetrahymena pyriformis*) taken from the literature (Cronin et al. 2000). The complete data set comprises 140 compounds, two *X*-variables [log *P* and energy of the lowest unoccupied molecular orbital (LUMO)], and one *Y*-variable (the endpoint). The two *X*-variables are almost uncorrelated with a squared correlation coefficient of 0.044. The data set is known to contain five outliers. The MLR results are summarized in Figure 3.

Figure 3A shows the relationship between observed and predicted endpoint data. This is a standard plot in QSAR analysis. In this case, a few outliers are identifiable, but sometimes the situation can be a bit trickier.

A diagnostic tool that is specifically designed to pinpoint outliers is the normal probability plot of residuals (Box et al. 1978). All observation points that lie on an imagined straight line that goes through the point (zero residual, 0.5 probability) have approximately normally distributed residuals. Any point that falls off such an imagined straight line has a residual, that is, a difference between measured and predicted endpoint data that is much larger or smaller than would be expected based on the assumption of nearly normally distributed residuals. A normal probability plot of the example data set is shown in Figure 3B. It is immediately evident from this plot that there are five outliers in the data set that all fall off the straight line.

After removal of the five outliers and refitting of the model, the normal probability plot looks much nicer (Figure 3C). We emphasize that deleting outliers should be done with caution so that the model is not overtrained. For the updated model (devoid of the five outliers) the explained *Y*-variation (R^2Y = 0.85) is 0.85 and the predicted *Y*-variation (Q^2Y = 0.84; estimated with cross-validation) is 0.84, which are excellent performance statistics. The regression coefficients are plotted in Figure 3D. We have chosen to plot them to simplify comparison with the PLS model (see Figure 4C). The regression equation is listed in the figure legend. As seen from the 95% confidence intervals, the uncertainty in these coefficients is very small.

Thus, in conclusion, we have a very good MLR model. Minor improvement in R^2Y and Q^2Y (2% in each parameter) is accomplished if the cross-term log *P**LUMO is included in the model; however, the significance of this slight improvement remains unclear.

Interpretation with a bit wider scope: defining the range of the QSAR model. It is possible to calculate the applicability domain of a

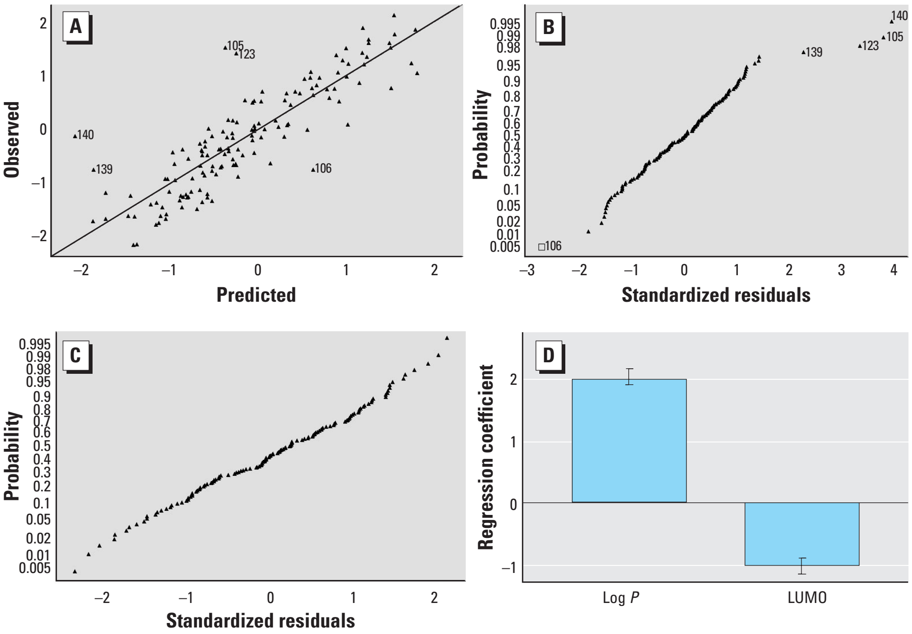

Figure 3. (*A*) Relationship between observed and predicted endpoint data. The five outliers are enumerated. (*B*) Normal probability plot of *Y*-residuals. Points falling off the dominant straightline structure are outliers. These have larger difference between observed and predicted y-data than would be anticipated assuming the *Y*-residuals to be (nearly) normally distributed. (*C*) Normal probability of *Y*-residuals of revised QSARs. There are no outliers in the revised model. (*D*) Regression coefficients of scaled and centered variables. The full regression equation is log EC_{50} = 0.2(±0.04) + 2.0(±0.08) log *P* – 1.0 (±0.06) LUMO + *e*. The error bars represent 95% confidence interval.

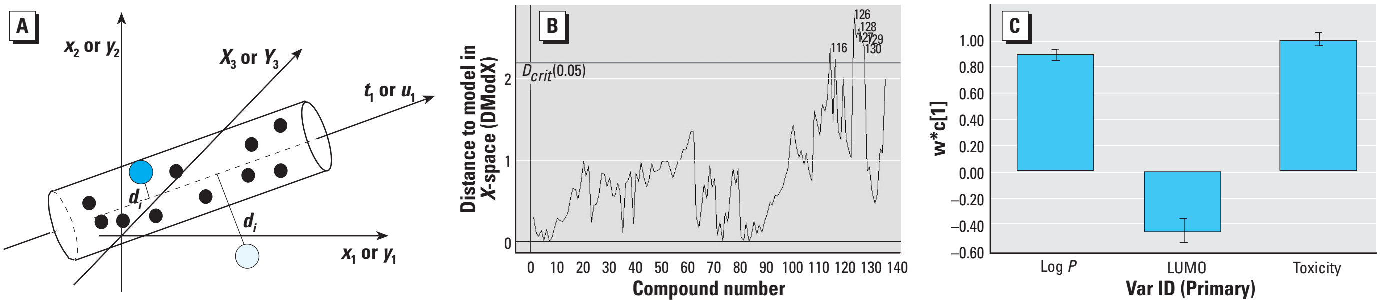

Figure 4. (*A*) A geometrical interpretation of a compound's distance to the model (DModX). A value for DModX can be calculated for each compound and these values may be plotted together with a typical deviation distance (D_{crit}) in order to reveal moderate outliers. D_{crit} can be interpreted as the radius of the "beer-can" inserted around the compounds. d_i, DModX of compound *i*. (*B*) Plot of DModX for the compounds of the first example. There are six moderate outliers in the *X*-data, that is, compounds 116 and 126–130. Analagous to *A*, these compounds can be understood as being positioned outside the "beer can" outline of the model. (*C*) PLS loadings of the first example. These show that log *P* is more important than LUMO, and the former parameter is approximately twice as large as the latter.

QSAR model, that is, the range within which it "tolerates" a new molecule. This specification can be made regarding both the *X*- and the *Y*-data as long as not all initial variance is used in the model. We will now illustrate this possibility.

Reanalyzing the above example with PLS yields a similar model with $R^2Y = 0.84$ and $Q^2Y = 0.84$ (Figure 4A–C). However, the PLS model uses only 42% [$R^2X = 1 - RSS_X/SS_X$ (residual sum of squares/sum of squares; "explained *X*-variation") = 0.42] of *X* to explain and predict the *Y*-data, not 100% as does MLR. The *X*-residuals are of diagnostic interest. They can be used to calculate the typical distance to the model in the *X*-data (here abbreviated DModX) for a compound (Figure 4A). Figure 4B shows DModX for each compound. We can also see the critical distance corresponding to the 0.05 probability level. This critical distance indicates the "tolerance volume" around the model, that is, the range of the model in the *X*-data (Eriksson et al. 2001). Apparently, a few compounds are positioned outside the range of the model; that is, they do not fit the model well. Hence, predictions for any of these should be considered with caution. Finally, Figure 4C provides the PLS loadings, which are reminiscent of the MLR regression coefficients (compare with Figure 3D).

A companion parameter to DModX, DModY, is also calculable. DModY is especially useful in the situation when more than one *Y*-variable is modeled by the same QSAR. This will be illustrated by our second example, where QSAR modeling is attempted for a set of 15 mononitrobenzene derivatives (Eriksson et al. 1995). The goal in this study was to be able to model and predict the aquatic toxicity profiles of the 15 chemicals based on information concerning their chemical properties. The 15 compounds were characterized using eight descriptor variables [boiling point (Bp), melting point (Mp), density (eta), log *P*, σ^-, HOMO (energy of the highest occupied molecular orbital), LUMO, and hardness). In total, eight biological responses were available (Deneer et al. 1987, 1989). These are primarily related to toxicity toward the four aquatic species *Poecilia reticulata*, *Daphnia magna*, *Chlorella pyrenoidosa*, and *Photobacterium phosphoreum*.

The data analysis resulted in a QSAR with $R^2X = 0.84$, $R^2Y = 0.76$, and $Q^2Y = 0.67$, which are excellent performance statistics considering that eight responses are handled simultaneously. For the interpretation of this QSAR model, we may consider the model coefficients (scores and loadings) to see how the compounds and the *X*- and *Y*-variables are interrelated (Figures 5A, B).

Figure 5A indicates that all *X*-variables load strongly in the model, and that *D*, Mp, σ^-, and LUMO are closely related. A second group is formed by log *P*, Bp, and η, whereas HOMO provides information different from these two groups. Overall, log *P* is the most important *X*-variable. Altogether, nitrobenzene (Figure 5A–C, point 1) is the least toxic compound to these aquatic organisms, and it is also the least hydrophobic compound (lowest value of log *P*).

Figure 5B shows the model scores. There are no outliers in the score space because all compounds lie inside the elliptic 95% tolerance volume depicted in the plot. This tolerance volume is given by a diagnostic called Hotelling's T^2. Hotelling's T^2 is a multivariate generalization of Student's *t*-test. It provides a check for compounds adhering to multivariate normality (Jackson 1991).

Plots of DModX and DModY are given in Figure 5C and D. These parameters suggest that this data set contains no outliers, neither in the *X*- or the *Y*-data. This absence of outliers is a valuable indication about chemical similarity and biological homogeneity among the studied compounds.

Thus, as shown here, there are two complementary diagnostic tools available, Hotelling's T^2 and DModX/DModY, that jointly assess the range of a QSAR model. The difference lies in the fact that, whereas DModX and DModY are derived from the unexplained X- and Y-variances (residuals), Hotelling's T^2 is founded with the explained variances. Further, through these diagnostics it is also possible to discriminate between strong (Hotelling's T^2) and moderate (DModX/DModY) outliers, depending on which tool is used for their detection.

A similar way of defining the range of a QSAR model is according to the leverage of a compound. The leverage *h* (Atkinson 1985) of a compound measures its influence on the model. It is noted that the leverage *h* and Hotelling's T^2 are, apart from a proportionality constant, identical. The leverage of a compound in the original variable space is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, \ldots, n), \qquad [2]$$

where x_i is the descriptor vector of the considered compound and *X* is the model matrix derived from the training set descriptor values. The warning leverage h^* is defined as follows:

$$\bar{h}^* = 3 \times h = 3 \times \Sigma_i\, h_i/n = 3 \times p'/n \quad (i = 1, \ldots, n), \qquad [3]$$

where *n* is the number of training compounds and p' is the number of model parameters.

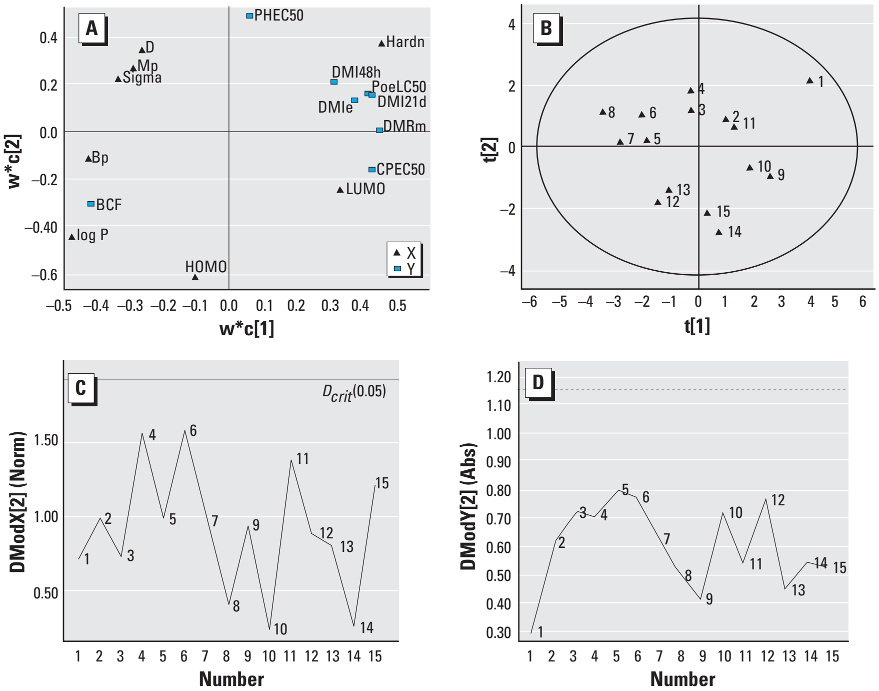

Figure 5. (*A*) PLS loading plot of the second data set showing the relationships between the eight *X*- and the eight *Y*-variables (boxed) at the same time. In the model interpretation one considers the distance to the plot origin. The farther away from the plot origin an *X*- or *Y*-variable lies, the stronger the model impact that particular variable has. In addition, we must also consider the sign of the PLS loading, which informs about the correlation among the variables. For instance, the *X*-variable log *P* is influential for the *Y*-variable BCF. This is inferred from their closeness in the loading plot. Hence, when log *P* increases BCF increases. In a similar way, we can see that HOMO is the most influential descriptor regarding the response PHEC50, albeit with an inverse relationship. For a description of the *Y*-variables, see Ericksson et al. (1995). (*B*) Distribution of the 15 compounds in the latent variable space, defined by the scores of the first two latent variables. The ellipse indicates the model applicability domain as defined by Hotelling's T^2. (*C*) DModX of the second data set. There are no moderate outliers in the *X*-data; that is, all compounds are inside the model tolerance volume. DModX is a way of defining the applicability domain of the QSAR in the *X*-space. (*D*) DModY of the second data set. There are no moderate outliers in the *Y*-data; that is, all compounds are inside the model tolerance volume. DModY is a way of defining the applicability domain of the QSAR in the *Y*-space.

Leverage values can be calculated for both training compounds and new compounds. In the first case, they are useful for finding training compounds that influence model parameters to a marked extent, resulting in an unstable model. In the second case, they are useful for checking the applicability domain of the model. A leverage greater than the warning leverage h^* means that the compound predicted response can be extrapolated from the model, and therefore, the predicted value must be used with great care. Only predicted data for chemicals belonging to the chemical domain of the training set should be proposed. The kind of leverage plot seen in Figure 6 allows a graphical detection of both the outliers and the influential chemicals in a model.

Yet another way of defining the range of a PM is according to the principles of optimum prediction space (OPS) by Gombar (1996). Although OPS has similar scope and objective as the combination Hotelling's T^2/DModX, the implementation is somewhat different. OPS is defined in the original variable space, whereas Hotelling's T^2 is usually based on calculation in score space of latent variable projection methods (Eriksson et al. 2001; Jackson 1991).

Assessing Predictive Power

Realizing the difference between fit and predictive power. In any modeling, including QSAR modeling, it is easy to manipulate data such that an apparently good model can be formulated. The most drastic step here is removal of observations (compounds) and variables that "do not fit" according to some subjective criterion. Furthermore, variables might be unduly transformed, and model complexity might be driven beyond pertinent limits. Such an inappropriate model often arises when one is merely interested in the fit of the model to the underlying data, and neglects its performance with new compounds. The problem with this kind of model is that it is not representative for other, additional compounds. Predictive validation is one way to reliably assess model adequacy for new compounds.

In this context, it is of crucial importance to realize the difference between a model's fit and prediction ability. The fit, usually estimated as R^2Y, tells how well we are able to mathematically reproduce the endpoint data of the training set. The problem with the goodness of fit is that with sufficiently many free parameters in the model, R^2Y can be made arbitrarily close to the optimal value of 1.0. Fortunately, the prediction ability is not as easy to manipulate. It measures how accurately we can predict the data of new compounds not previously used in the model training. The predictive power of a model may be estimated by the goodness of prediction parameter Q^2Y.

The most demanding way to predictively validate a model is by external validation, which consists of making predictions for an independent set of data not used in the model calibration. Such a prediction set may well be selected according to the principles of multivariate design. However, external validation might not always be tractable, for example, in QSARs because of the resources needed to make a test set of new compounds. Hence, alternatives for predictive validation are of interest, for example, methods such as cross-validation and permutation testing (Eriksson et al. 1997).

Cross-validation. In the two examples above, no external validation set was available, so we used cross-validation with seven cross-validation groups instead. Basically, cross-validation is performed by dividing the data in a number of groups and then developing a number of parallel models from reduced data with one of the groups deleted. It should be noted that increasing the number of cross-validation groups to N (number of compounds, that is, the so-called leave-one-out (LOO) approach, is not recommended because the estimated Q^2Y then becomes too similar to R^2Y (Eriksson et al. 2001; Shao 1993;).

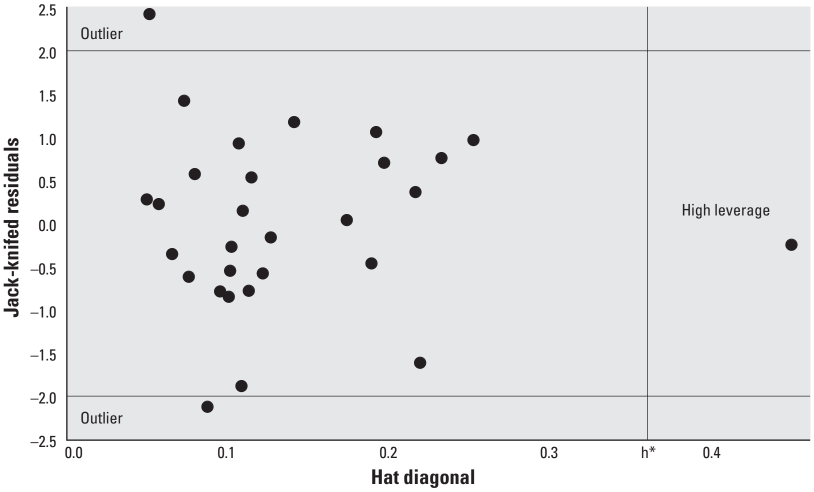

Figure 6. MLR outlier and leverage plot: outlier chemicals are points with jack-knifed (cross-validated standardized) residuals greater than two standard deviation units; influential chemicals are points with high leverage values higher than the warning value h^*.

After developing the reduced model, the omitted data are used as a test set, and the differences between actual and predicted Y-values are calculated for these data points. The sum of squares SS of these differences from all the parallel models are used to form PRESS (predictive residual sum of squares). This is a measure of the predictive ability of the model and is often reexpressed as Q^2Y (the "cross-validated" R^2Y), a statistic that is similar to R^2Y.

Without a high R^2Y, it is impossible to obtain a high Q^2Y. Generally, a $Q^2Y > 0.5$ is regarded as good and a $Q^2Y > 0.9$ as excellent, but these guidelines are of course heavily application dependent (Eriksson et al. 2001). Differences between R^2Y and Q^2Y larger than 0.2–0.3 indicate the presence of many irrelevant model terms or a few outlying data points.

We note that the R^2Y and Q^2Y measures can be equivalently expressed as residual standard deviations and predictive residual standard deviations (PRESDs). The latter is often called SDEP (standard error of prediction). These standard deviations should be of sizes similar to those of the known or expected "noise" in the system, for example, ±0.3 units for log(1/C) in QSAR investigations.

Response permutation testing. One limitation of cross-validation is that it assesses only the predictive power and provides no statement of the statistical significance of the estimated predicted power. To obtain an estimate of the significance of a Q^2Y value, one may develop a number of parallel models based on fit to randomly reordered Y-data, and then evaluate the real Q^2Y in light of a distribution of Q^2Y values of reordered response data. A good description of permutation testing can be found in Van der Voet (1994).

This validation option works as follows: For the training set, the X-data are left intact, whereas the Y-data are permuted to appear in a different order. This means that the Y-data remain numerically the same, but their positions are shifted by random shuffling. A QSAR model is then fitted to the permuted Y-data, and by using cross-validation, both R^2Y and Q^2Y values are computed for the derived model. These "permuted" values may then be compared with the estimates of R^2Y and Q^2Y of the "real" model to get a first indication of the significance of the latter values.

In the next round, a second model is fitted to another permuted version of the Y-data, and new estimates of "permuted" R^2Y and Q^2Y values are thus formed. By repeating this permutation procedure a number of times, say, between 50 and 100 times, and by establishing an equivalent number of parallel QSAR models, it is possible to achieve reference distributions of

R^2Y and Q^2Y based on random data. Such reference distributions are useful for appraising the statistical significance of the R^2Y and Q^2Y parameters of the parent QSAR model. If the "real" Q^2Y and R^2Y are found outside such reference distributions, this constitutes a strong indication of a valid model.

Furthermore, because the numerical values of the "permuted" versions of R^2Y and Q^2Y depend, at least partly, on the extent of perturbation inflicted by the permutation procedure, it is advisable to keep track of the correlation coefficient between original and permuted *Y*-variables. Should an original *Y*-variable be only mildly perturbed by permutation, the permuted *Y*-variable will by necessity display a high correlation coefficient with the original *Y*-variable. By jointly assessing such correlation coefficients and "permuted" R^2Y/Q^2Y numbers, it is possible to understand and explain the existence of occasionally high R^2Y and Q^2Y values for permuted *Y*-data.

An informative way of summarizing results of response permutation testing was recently published (Eriksson et al. 2001). Figure 7 shows such a plot for the first data set. It manifests the validity of that QSAR because the "real" model parameters are constantly much higher than their permuted counterparts. The plot in Figure 7 was constructed by letting the *y*-axis represent the R^2Y/Q^2Y values of all MLR models, including the "real" one and by assigning the *x*-axis to the correlation coefficients between permuted and original response variables. Observe that the points of R^2Y and Q^2Y for the original model are always found in the right-hand part of the plot at correlation 1.0 (because 1.0 is the correlation coefficient obtained when correlating a variable with itself).

Assessing Parameter Uncertainty

Confidence intervals. When estimating a parameter, for example, a regression coefficient, we would like to know the significance of this parameter—we would like to know not only the estimated value of the statistic but also how precise it is. In other words, we want to be able to state some reference limits within which we may reasonably declare the true value of the statistic lies. Such statements may assert that the true value is unlikely to exceed some upper limit, or it is unlikely to be less than some lower limit, or it is unlikely to lie outside a pair of limits. Such a pair of limits is often known as confidence limits or a confidence interval and is just as important as the estimated statistic itself. The degree of confidence used is usually set at 95%, but higher or lower levels may be chosen by the user.

Usually, in QSAR modeling parameter uncertainty is given in terms of 95% confidence intervals. Such intervals are easily calculated in MLR when applied to well-conditioned data sets. For other methods and more challenging data sets, more elaborate calculations are often necessary (Burnham et al. 1996, 1999, 2001; Denham 1997).

Jack-knifing. One way to estimate standard errors and confidence intervals directly from the data is to use jack-knifing (Efron 1982; Efron and Gong 1983). This is useful for data where the assumptions of regression analysis are not fulfilled. The objective of jack-knifing is to estimate variability of model parameters.

Interestingly, cross-validation—where the objective is to estimate the model complexity giving the optimal predictive power—produces results that can be fed directly to jack-knifing. In this way, the various submodels generated by cross-validation are used to calculate the standard errors of the model parameters, which are then converted into confidence intervals via the *t*-distribution. This connection between cross-validation and jack-knifing was highlighted by Herman Wold in 1982 and has recently been revived in Martens and Martens (2000).

Bootstrapping. Another way to estimate confidence intervals for model parameters is to use the method of bootstrap resampling. The basic premise of this method is that the data set is representative of the population from which it was drawn. Because there is only one data set, bootstrapping simulates what would happen if the population were resampled by randomly resampling the data set (Efron and Tibshirani 1993; Wehrens et al. 2000). An illustration of the use of bootstrap resampling to derive confidence intervals for the parameters of a classification model is provided in Worth and Cronin (2000).

Variable Selection and Reduction

A delicate problem. Yet another approach to improving QSARs is the deletion of uninformative variables. However, one should be very careful when reducing the number of variables because this can be done in so many ways, so almost any result is possible. In variable selection, it is important to test the predictive power of the model on real new data, and not just the cross-validation with the training set. In multivariate data, most of the *X*-variables contain at least some information about *Y*. Hence, one can hope for only a mild variable reduction, usually not more than 20–30% of the variables have less information than the noise level (Eriksson et al. 2001).

Moreover, because of the correlations among the more important *X*-variables, one can continue to reduce the *X*-variables further than these 20–30% with no apparent decrease in fit. This makes the remaining *X*-variables take over importance from the ones that are deleted, and a serious bias is introduced. Thus, the interpretation of the model shifts, and some variables take the role of being related to *Y*, while other variables correlated to these have been deleted and hence are forgotten in the interpretation. This also makes the prediction power of the model deteriorate because the correlations are not perfectly stable, and for new samples/molecules, important variables are now missing in the model.

Also, one should remember that even seemingly unimportant variables still have a role in diagnosing outliers. Consider a variable that is almost constant in the training set and that will appear unimportant in the QSAR model. If a new compound has a value of this variable that substantially differs from its values in the training set, this is an indication that this compound is different, and hence predictions of its *Y*-values are doubtful. If one mechanically deletes all variables that do not contribute to the modeling of *Y* in the training set, one automatically decreases the possibility of finding outliers among the new observations.

How do we then handle very many variables? Despite all the drawbacks discussed above, and the care that should be taken in selecting variables, variable selection is sometimes necessary to find a simple and predictive

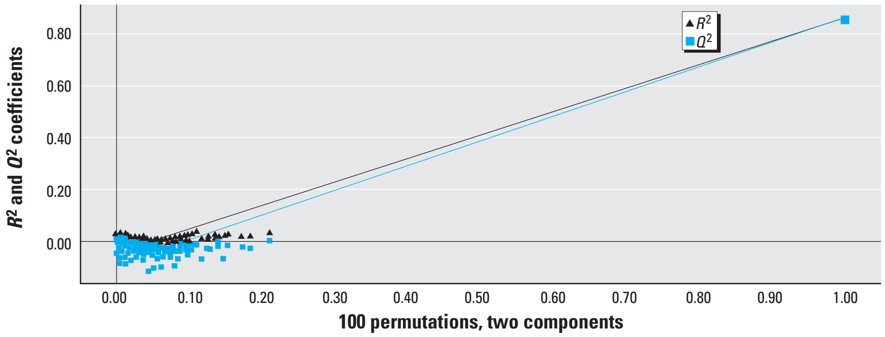

Figure 7. Response permutation plot of the first example. The *y*-axis represents R^2Y (triangles) and Q^2Y (squares) for every model, and the *x*-axis designates the correlation coefficient between original and permuted response data. One way of summarizing this plot is to conduct conventional regression analysis in the two sets of points; that is, one regression line is fitted among the R^2Y points (triangles) and another line among the Q^2Y points (squares). The intercepts of the resulting regression lines are interpretable as measures of "background" R^2Y and Q^2Y obtainable with zero correlation between original and permuted data. Experience shows that the R^2Y intercept should not exceed 0.3 and that the Q^2Y intercept should not exceed 0.05. Intercepts below these limits indicate valid models. Toxicity intercepts: R^2 =0.0, –0.0388, Q^2 = 0.0, –0.0894.

QSAR model. Nowadays it is becoming quite common to use a wide set of molecular descriptors of different kinds (experimental and/or theoretical) able to capture all the structural information possibly related to the Y-response. A recent survey of this was published by Livingstone (2000). Many software programs calculate wide sets of different theoretical descriptors, from SMILES (simplified molecular input line entry specification), two-dimensional graphs, and three-dimensional *x,y,z*-coordinates. Only some of the more complete are mentioned here: ADAPT (Jurs 2002; Stuper and Jurs 1976), OASIS (Mekenyan and Bonchev 1986), CODESSA (Katritzky et al. 1994), and DRAGON (Todeschini et al. 2001). It has been estimated that more than 3,000 molecular descriptors are now available, and most of them are summarized and explained in recently published books (Devillers and Balaban 1999; Karelson 2000; Todeschini and Consonni 2000). The great advantage of theoretical descriptors is that they are calculable for not yet synthesized chemicals.

There are two main steps in QSAR modeling by variable selection: first, statistically validated and robust regression models must be found, and second, the model variables must be interpretable. In principle, all the different possible variable combinations of the *X*-variables should be investigated to find the most predictive QSAR model. However, this may be quite taxing, mainly for reasons of time. Thus, first, various types of rapid prescreens (discarding constant values, pair-correlated variables, etc.) are often implemented to sort out a limited set of descriptors among which the selection of those really related to the response, not only in fitting but most importantly in prediction, is then performed by alternative variable selection methods.

Several strategies for variable subset selection have been applied in QSARs (among those most widely applied: stepwise regressions, forward selection, backward elimination, simulated annealing, and evolutionary and genetic algorithms). A recent comparison (Xu and Zhang 2001) of these methods has given a demonstration of the advantages and success of genetic algorithms as a variable selection procedure for QSAR studies. Below, we discuss genetic algorithms and a few alternatives.

Genetic algorithm strategy for variable selection. Genetic algorithms are a particular kind of evolutionary algorithm shown to be able to solve complex optimization problems in a number of fields, including chemistry (Davis 1991; Goldberg 1989; Hibbert 1993; Wehrens and Buydens 1998). The natural principles of the evolution of species in the biological world are applied: the assumption that conditions that lead to better results will prevail over poorer ones, and that improvement can be obtained by different kinds of recombination of independent variables, that is, reproduction, mutation, and crossover. The goodness of the selected solution is measured by a response function that has to be optimized.

Genetic algorithms, first proposed as a strategy for variable subset selection in multivariate analysis by Leardi et al. (1992), are now widely and successfully applied in QSAR approaches where there are many molecular descriptors as *X*-variables in various modified versions, depending on the way to perform reproduction, crossover, mutation, and so forth (Devillers 1996; GFA of Rogers and Hopfinger 1994; Leardi 1994; MUSEUM of Kubinyi 1994a, 1994b; MOBY-DIGS of Todeschini 1997).

In variable selection for QSAR studies, each variable (molecular descriptor) is denoted by a bit equal to 1 if present in the regression model or to 0 if excluded. A population constituted by a number of 0/1 bit strings (each of length equal to the total number of variables) is evolved following genetic algorithm rules, maximizing the predictive power of the models (explained variance in prediction, Q^2Y, or root mean squared error of prediction). Only the models producing the highest predictive power are finally retained and further analyzed.

Whereas revolutionary algorithms search for the global optimum and end up with only one or very few results (Kubinyi 1994a, 1994b, 1996), genetic algorithms simultaneously create many different results of comparable quality in larger populations of models. Within a given population, the selected models can differ in number and kind of variables.

Different rules can be adopted to select the final "best" models. Todeschini, Gramatica, and colleagues (Gramatica et al. 1998, 1999, 2000; Gramatica and Papa 2003; Todeschini and Gramatica 1997) use the QUIK rule (*Q* under influence of *K*) (Todeschini et al. 1999) to avoid multicollinearity without prediction power or "apparent" prediction power (chance correlation). According to this rule, only models with a *K* multivariate correlation calculated on the *X* + *Y*-block that is at least 5% greater than the *K* correlation of the *X*-block are considered statistically significant. Alternatively, one may use the approach of Hopfinger (discussed in a later section).

Model validation is always used to avoid "overfitted" models, that is, models where too many variables have been selected, and to avoid selecting variables randomly correlated with the dependent response. Particular care must be taken against overfitting; therefore, subsets with fewer variables are favored, even though the chance of finding "acceptable" models increases with increasing the selected variables. The proportion of random variables selected by chance correlation could also increase (Jouan-Rimbaud et al. 1996).

The collinearity in the original set of molecular descriptors results in many similar models yielding more or less the same predictive power. Therefore, after having selected a set of similar PMs, model validation proceeds via leave-more-out cross-validation, response permutation testing (Y-scrambling), bootstrapping (Efron 1982), or other resampling techniques. This is done to avoid overestimation of the model predictive power by Q^2_{LOO} (Golbraikh and Tropsha 2002; Shao 1993), to verify model predictivity stability, and to select the "best" model. Finally, for the strongest evaluation of model applicability for prediction in new chemicals, external validation (verified by Q^2_{EXT}) of all the models is also recommended, depending on whether the data set is large enough to permit an independent external validation set. The best splitting of the original data set into a representative training set and a validation set can be obtained by applying experimental design (Eriksson et al. 2000b; Marengo and Todeschini 1992).

If after several different runs of genetic algorithms the same subsets of variables have been selected, and if the obtained models pass all the validation procedures above (cross-validation), external testing, *Y*-scrambling, bootstrapping), there is a reasonable certainty that the models are robust and applicable for prediction. Good predictive properties is also an indication that chance correlation has been avoided.

Because genetic algorithms simultaneously create many different good models in a population, the user can choose the "best model" according to need: the interpretability of the selected molecular descriptors, the possibility of having reliable predictions for some chemicals rather than others, the highlighting of different outliers, and so forth. The need for interpretability depends on the application, as a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the 'mechanism' in chemical terms, but it is often not necessary, per se (Livingstone 2000). This type of QSAR model follows a path that starts with a statistical validation and further interpretation for their biological and mechanistic meaning (Tropsha et al. 2003). Therefore, their application domain is mainly related to the production of predicted data, verified for their reliability.

Assessing model uniqueness. Hopfinger and colleagues advocate a related approach aiming at defining the best QSAR model. This approach is based on some of the elements described above, notably, cross-validation, response permutation testing, and variable selection (Kulkarni et al. 2001). They strive to maximize Q^2Y through elimination of unimportant *X*-variables. Several different model versions are derived using genetic algorithms and the ones producing the highest Q^2Y are retained. In the next step cross-correlation analysis of the modeling residuals from the set of best models is used to determine how many unique models have been obtained. A unique model will have low correlations of its residuals of fit to those of the alternative top-ranked models.

After having selected a set of unique models with highest possible Q^2Y, model validation proceeds via response permutation testing and/or external predictive validation, depending on whether the data set is large enough to permit an independent external prediction set. In some cases when there is thought to be considerable noise in the *Y*-data, the approach of Hopfinger and colleagues also involves studying the stability of the resultant QSAR models as a function of increasing simulated error among the *X*-variables. The objective with this latter exercise is to investigate whether stable QSARs with respect to the inherent error of the data set have been obtained (Hopfinger AJ and Jaworska J. Personal communication).

GOLPE (generating optimal linear PLS estimations). About a decade ago an advanced variable selection procedure called GOLPE was introduced by Sergio Clementi and colleagues (Baroni et al. 1993) and has found widespread use in three-dimensional QSARs. The objective of this approach is to obtain PLS regression models with the highest prediction ability. The key steps of this approach involve a first preliminary variable selection by means of a *D*-optimal (determinant optimal) design in the loading space, and an iterative evaluation of the effects of the individual variables on the model predictivity. This is accomplished based on the validation of a number of partial submodels using many combinations of the descriptor variables as dictated by a fractional factorial design strategy. Cruciani and Watson (1994) show the utility of GOLPE in generating three-dimensional QSAR models with good predictive power.

Hierarchical modeling for easier model interpretation and as an alternative to variable selection. In two- and three-dimensional QSAR modeling involving many variables, plots and lists of coefficients, loadings, and so forth, rapidly become messy, and results are therefore difficult to interpret. As discussed above, there may then be a strong temptation to eliminate variables to obtain a smaller data set. Such a reduction of variables, however, often removes information and makes the modeling efforts less reliable. Model interpretation may be misleading, and predictive power may deteriorate.

As reported by Berglund et al. (1997), an interesting alternative is to partition the variables into blocks of logically related variables and apply hierarchical data analysis. All such blocks may be analyzed individually. This modeling forms the base level of the hierarchical modeling setup (Eriksson et al. 2002). The score vectors, often called "super variables," formed on the base level may be concatenated in new matrices amenable for analysis on the top level. On the top level, superficial relationships between the *X*- and the *Y*-data are investigated. On the base level, in-depth information is extracted for the different blocks.

Bayesian Methods for Reliability Testing

Bayesian-based methods have been heavily used in reliability engineering and diagnostic medicine where models are used for decision making. These methods are perfectly suitable to evaluating QSARs and have been introduced to the field but still are not used broadly (McDowell and Jaworska 2002; Pet-Edwards et al. 1989). One characteristic of Bayesian-based procedures is that they allow both prior information (including expert judgment) and sampling information to be combined in the weighting scheme inherent in Bayes' formula. The second characteristic of Bayesian-based methods is they can be formulated in a recursive form. This means Bayesian methods allow successive updating of battery interpretation as additional tests results are obtained, which is particularly useful if sequential testing procedures are being considered.

The most common and simple application of Bayes' approach is found in evaluating performance statistics for two-way categorical classifications. It uses as inputs sensitivity and specificity. Sensitivity is the fraction of active chemicals that are predicted to be active by the model (α_i^+); and specificity is defined as the fraction of nonactive chemicals the model predicts nonactive (α_i^-). Sensitivity can also be expressed as Pr(*P*+|*S*+), the conditional probability a model predicts a chemical to be active (*P*+) given that the true state is active (*S*+). Similarly, specificity is defined as Pr(*P*–|*S*–), the conditional probability the model predicts a chemical nonactive (*P*–) given the true state is nonactive (*S*–).

We then can use Bayes' formula

$$\Pr\left(S_i \middle| T_j\right) = \frac{\Pr\left(S_i\right)\Pr\left(T_j \middle| S_i\right)}{\sum_i \Pr\left(S_i\right)\Pr\left(T_j \middle| S_i\right)} \qquad [4]$$

to obtain $\Pr(S_i|T_j)$, the posterior probability of condition S_i prevailing given we have test result *j* from *a*) the prior probability of S_i, $\Pr(S_i)$, and *b*) $\Pr(T_j|S_i)$, the likelihood of *j*th test result given true state is S_i.

It is important to note the likelihood value (sensitivity or specificity), $\Pr(T_j|S_i)$, is conditional on S_i (not known to the observer or analyst), whereas the posterior probability is conditional on the *observed* result or prediction T_j. The posterior probability or predictive value is the appropriate statistic for inferring from test results the probability the modeled chemical has condition S_i. Posterior probabilities are statistically precise statements of the likelihood a chemical has a particular state or attribute, conditional on the test evidence obtained.

The predictive value positive (PVP), Pr(*S*+|*T*+), denotes the probability a chemical is active (*S*+) given a model predicts the chemical to be active. Predictive value negative (PVN), Pr(*S*–|*T*–), is the probability a chemical lacks the attribute, for example, is *S*–, given a negative model result is obtained. The terms sensitivity, specificity, PVP, and PVN are sometimes referred to as the Cooper statistics (Cooper et al. 1979). Confidence intervals for the Cooper statistics can be derived by bootstrap resampling (Worth and Cronin 2001a). The computational form for the two-way classification problem using Bayesian revision is presented in Table 1.

This approach can easily be extended to an *n*-way classification problem. This analytical framework is easily extended to a system with *n* possible states or characteristics. Extending this to *n* possible states requires the same parameters used in the two-state analysis but describing *n* possible states and *n* possible predictions: *a*) prior probabilities for each possible state, $\Pr(S_i)$, for $i = 1, 2, \ldots n$; and *b*) likelihood values, $\Pr(P_j|S_i)$, where P_j is a model prediction of the *j*th state given true state is S_i. These likelihood values form an $n \times n$ contingency table analogous to the two-state model's 2 × 2 contingency table of likelihood values

Table 1. Bayesian revision of diagnostic test result.

Prior probability		Likelihood		Joint probabilities: prior × likelihood		Posterior probabilities	
		$\Pr(T_j \mid S_i)$		$\Pr(S_i) \times \Pr(T_j \mid S_i)$		$\Pr(S_i \mid P_j)$	
S_i	$\Pr(S_i)$	*T*+	*T*–	*T*+	*T*–	*P*+	*P*–
S+	p	sens	1 – sens	$p \times \text{sens}$	$p \times (1 - \text{sens})$	$\frac{p \times \text{sens}}{p \times \text{sens} + (1 - p)(1 - \text{spec})}$	$\frac{p \times \text{sens}}{p \times (1 - \text{sens}) + (1 - p) \times \text{spec}}$
S–	$1 - p$	1 – spec	spec	$(1 - \text{p}) \times (1 - \text{spec})$	$(1 - \text{p}) \times \text{spec}$	$\frac{(1 - p) \times (1 - \text{spec})}{p \times \text{sens} + (1 - p) \times (1 - \text{spec})}$	$\frac{(1 - p) \times \text{spec}}{p \times (1 - \text{sens}) + (1 - p) \times \text{spec}}$
Sums	1.0			$p \times \text{sens} + (1 - p)(1 - \text{spec})$	$p \times (1 - \text{sens}) + (1 - p) \times \text{spec}$	1.0	1.0

comprised sensitivity, specificity, and their complements.

The application of Bayes' formula to an *n*-state model is identical to the two-state case (Equation 4) and the same conditions that hold for the two-state case apply to the *n*-state case:

$$\sum_i \Pr(S_i) = 1.0, \sum_i \Pr(\mathrm{P} \mid S_i) = 1.0 \qquad [5]$$

The PVP and PVN are not constant but vary with prior probability, Pr(*S*+) or Pr(*S*–). In other words, given fixed sensitivity and specificity, PVP and PVN vary according to the prevalence or proportion of active (toxic) chemicals in a population. This means the predictive capacities of QSAR models should not be judged according to these statistics alone because the investigator can give PVP and PVN almost any values by altering the prevalence of *S*+ in the test set. Examining these predictive statistics reveals the importance of evaluating the prior for understanding classification probabilities, correct and incorrect (Figure 8). As sensitivity and specificity increase, the probability curves become increasingly nonlinear. The prior probability of active/not active is not a property of an individual chemical; it is the relative frequency of active/not active in a population of chemicals. A chemical is either active or nonactive. It has no probability of being active or not (in a given biological test system under defined exposure conditions)—the probability we call prior in this case is a measure of uncertainty to its true state.

Sequential Use of Models

Sequential testing (QSAR testing) of all chemicals with models of even modest performance characteristics can significantly reduce misclassification rates when compared with single tests or multiple tests where only initial positives are subjected to subsequent tests. In most testing schemes, medical and otherwise, for economic reasons those testing negative to the first screening test are not subjected to confirmatory testing, provided that the screening test has a high sensitivity and therefore low probability of false negative test results among *S*+ objects. The actual rate or proportion of items testing false negative is not 1 – sensitivity but Pr(*S*+)Pr(*P*–|*S*+). It should be noted this applies in tiered assessment strategies for genotoxicity, but the converse is true in the tiered assessment strategies for skin and eye irritation, where negative findings *in vitro* are subject to confirmatory testing in animals (OECD 1998). Using QSAR models to classify chemicals does not have this limitation; once the QSAR model is developed, the marginal cost of making a QSAR "test" of a chemical is nearly zero. Thus, analysts using QSAR models have the opportunity to apply multiple tests to all chemicals without financial penalty. This allows great improvment in the reliability of predictions.

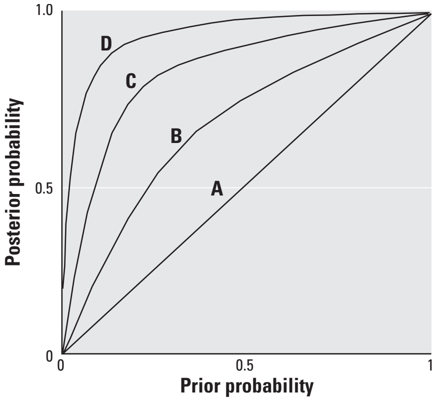

Figure 8. Range of posterior sequential predictions as a function of prior probability with no tests (*A*) , one test (*B*), two tests (*C*), and three tests (*D*).

Interestingly, a sequential testing approach for uncovering potential estrogenic endocrine disruptors has recently been adopted by the U.S. Food and Drug Administration (FDA). It is a method based on four phases (Hong et al. 2002; Shi et al. 2002; Tong et al. 2002).

Battery selection method. The reliability of predictions can be enhanced by using information from more than one model. The battery selection method (Pet-Edwards et al. 1989) is a system for evaluating and selecting batteries of tests; it was originally applied to carcinogenicity prediction and is therefore known as the CPBS (carcinogenicity prediction battery selection) method. The two CPBS methodologies have two main objectives, *a*) to determine the reliability and predictive capability of a battery of tests that individually may give mixed results; and *b*) to develop a strategy to formulate and select optimally preferred batteries of tests—optimal in terms of collective performance, minimum testing time, or costs or a compromise of these attributes.

The CPBS approach is a collection of methods designed to aid in selecting and interpretation tests used for decision making. The CPBS method relies on Bayesian decision theory to support sequential nature of the testing, cluster analysis to determine dependencies among the various models used in the battery, multiple-objective decision making to aid finding the optimal solution using cost, time and performance criteria, and dynamic programming to optimize the search for the best test battery when a number of tests are available.

The CPBS method consists of *a*) preliminary data analysis to evaluate and summarize information for use in battery selection with special attention on dependencies among tests and Bayesian prediction, *b*) battery selection, and *c*) Bayesian prediction to interpret the results.

The following initial strategy is advised in forming a battery of tests:

- An odd number of tests should be used to make the most decisive package (i.e., to be able to apply the "positive majority" rule). The battery is considered positive for the property if the majority of the results are positive for the property.
- If models with high sensitivity and specificity (both > 0.75) are available and statistically independent, use as many as is cost-effective.
- A model with high sensitivity (> 0.75) and lower specificity (< 0.75) should always be coupled with a model with high specificity (> 0.75) and lower sensitivity (< 0.75).
- Avoid models with low sensitivity and low specificity.

For the further refinement of this initial strategy, see Pet-Edwards et al. (1989). The majority, consensus, or probability limit criteria are used for selecting the best test battery. These decision rules need not guide the inference once a test battery has been selected and test results obtained; the posterior probability values—Pr(S_i|test results)—indicate the appropriate inference. The decision about how to act on this information is a more complicated question involving the consequences of each decision/outcome.

The reason for the third recommendation can be illustrated by considering a battery of just two tests, one with high sensitivity, the other with high specificity. The high-sensitivity test is used to detect the attribute of interest, such as the presence of a certain type of toxicity in a set of chemicals. This test correctly identifies most chemicals that exhibit the toxicity, but it does so at the expense of overpredicting the toxicity of chemicals that lack toxic potential; that is, it generates too many false positives. When such a test is combined with a high-specificity test, the latter test serves to confirm most correct positive predictions of the first test while correctly identifying most false positives from the first test (*S*–*P*+) as negative on the second test. The latter occurs because the second test correctly classifies most *S*– items by virtue of its high specificity, Pr(*P*–|*S*–).

In this paired arrangement of two tests, the high-sensitivity test is sometimes called the detection or screening test, whereas the high-specificity test is sometimes called the confirmation test (Feinstein 1975). A chemical would be predicted as negative (nontoxic) if the outcome of the detection was negative, whereas the chemical would only be predicted as positive (toxic) if the outcomes of both the detection and confirmation tests were positive. For QSARs the cost of making predictions is marginally low, so both positive and negative chemicals are tested through the whole battery. This is demonstrated in the example below.

Predictivity of independent tests. We now focus on the predictivity based on a battery of *k* tests. Let $\alpha_1^+, \alpha_2^+, \ldots, \alpha_k^+$ be the sensitivities of the tests and $\alpha_1^-, \alpha_2^-, \ldots, \alpha_k^-$ be the specificities of the tests.

For the case where the *i*th tests gave positive results, predictivity of the entire battery is calculated using the recursive formula

$$\Pr\left(S_i^+ \middle| P_i\right) = \frac{\Pr\left(S_{i-1}^+ \middle| P_{i-1}\right)\alpha_i^+}{\left(1-\alpha_i^-\right)+\left(\alpha_i^+ + \alpha_i^- - 1\right)\times \Pr\left(S_{i-1}^+ \middle| P_{i-1}\right)}. \qquad [6]$$

Similarly, for the case where the *i*th test gave negative results, the predictivity of the entire battery is given by

$$\Pr\left(S_i^+ \middle| P_i\right) = \frac{\Pr\left(S_{i-1}^+ \middle| P_{i-1}\right)\times\left(1-\alpha_i^+\right)}{\alpha_i^- - \left(\alpha_i^+ + \alpha_i^- - 1\right)\Pr\left(S_{i-1}^+ \middle| P_{i-1}\right)}. \qquad [7]$$

After each test there is a refinement in the predictivity. This can be visualized as in Figure 8, which shows the range of posterior sequential predictions as a function of prior probability with no tests, one test, two tests, and so forth.

The following example demonstrates a two-model sequential classification procedure using QSAR models to classify chemicals and has been previously described in McDowell and Jaworska (2002). Those authors assumed existence of two QSAR models to predict a particular chemical characteristic. The first test has sensitivity of 0.95 and specificity of 0.85. The values are reversed for the second test: sensitivity is 0.85 and specificity is 0.95. All chemicals are subjected to both tests. The resulting predictive values for all the tests are based on a prior probability of "active" of 0.10. The results are summarized in Table 2.

The columns of Table 2 denoted "Prior × likelihood" contain the relative frequency of classification rates for each test; the misclassification rates are summarized in columns 11 and 12. Total misclassification rate for model 1 is 14% in total; 13.5% is misclassified as false positive, 0.5% as false negative. When model 1 positives are subjected to model 2, 6.2% test false negative (column 5) and 2.9% test false positive (column 4), totaling 9.1% (column 9) misclassified. When adjusted to reflect the population proportion that predicted positive by model 1, the model 1 positives misclassified by model 2 represent 2.1% of total population with 1.4 and 0.7% testing false negative and false positive, respectively. Summing up the misclassification rates for model 2 shows a total misclassification rate of 6%, 1.5% as false negatives, and 4.5% as false positives. This represents a 57% decline in total misclassification rate compared with using one test. Note that this example does not rely on the majority rule introduced above. Rather, probabilities that a chemical is positive or negative conditional on the whole battery are explicitly calculated. Therefore, this approach can be used for two-model batteries and applies to any $n > 1$ test batteries.

Predictivity of dependent tests (Pet-Edwards et al. 1989). If the tests are dependent, a correction factor needs to be introduced expressed as conditional dependence:

$$K_P\left(r_1,r_2\right) = \frac{\Pr\left(r_1,r_2 \middle| P\right)}{\Pr\left(r_1 \middle| P\right)\Pr\left(r_2 \middle| NP\right)} \qquad [8]$$

and

$$K_{NP}\left(r_1,r_2\right) = \frac{\Pr\left(r_1,r_2 \middle| NP\right)}{\Pr\left(r_1 \middle| NP\right)\Pr\left(r_2 \middle| NP\right)} \qquad [9]$$

and the batch formula for predictivity is preferred over a recursive one:

$$\Pr\left(P \middle| r_1,r_2,\ldots,r_k\right) = \frac{1}{1+\dfrac{\Pr(NP)K_{NP}\left(r_1,\ldots r_k\right)\Pr\left(r_1 \middle| NP\right)\mathrm{L}\ \Pr\left(r_k \middle| NP\right)}{\Pr(P)K_P\left(r_1,\ldots r_k\right)\Pr\left(r_1 \middle| P\right)\mathrm{L}\ \Pr\left(r_k \middle| P\right)}} \qquad [10]$$

When r_i is positive, then $\Pr(r_i|P)$ would be sensitivity of test *i* and $\Pr(r_i|NP)$ would be 1 – specificity of test *i*. Similarly, if r_i was negative, then $\Pr(r_i|P)$ would be 1 – sensitivity of test *i* and $\Pr(r_i|P)$ would be specificity of test *i*.

Discussion

The Need for Reliability Assessment of QSAR and Related Models

Executive summary reports of two recent QSAR projects in the European Union (Anonymous 1995, 1999) indicate that there are many "environmental" QSAR models available. It is clear that these models cover broad classes of chemicals for many of the environmental endpoints that are used in the risk assessment of existing and virtual chemicals. A primary conclusion of these two projects, however, was that if such models are used for prediction outside their applicability domains, very unreliable predictions may result (Anonymous 1995, 1999). Consequently, in these reports, the reader is frequently reminded about the necessity of clearly defining the boundaries of each model. These reports also point out that it should be realized that our predictive capabilities are limited because for several classes of compounds or for very specific mechanisms of action, the QSAR models are simply not available and the progress in establishing such models is slow.

For both existing (published) and putative (still under development) QSAR models, it is important that reliability be assessed carefully and consistently. Reliability assessment procedures of QSARs must consider several aspects, for instance, quality of the underlying data, the chemical domain of the training set, predictivity estimates, and the work flow underpinning the QSAR. The predictions using any given QSAR model should be restricted to the chemicals that belong to the model domain. This emphasizes the importance of understanding the compositions of the

Table 2. Hypothetical application of sequential screening tests and associated misclassification rates.

Prior probability		Likelihood $\Pr(T_j \mid S_i)$		Prior × likelihood $\Pr(S_i)\Pr(T_j \mid S_i)$		Posterior probability $\Pr(S_i \mid T_j)$			Proportion misclassified	
S_i	$\Pr(S_i)$	T+	T–	T+	T–	T+	T–	Category	Individual test	In population
Test 1										
S+	0.100	0.950	0.050	0.095	0.005	0.413	0.006	FN	0.005	NA
S–	0.900	0.150	0.850	0.135	0.765	0.587	0.994	FP	0.135	NA
Totals	1.000			0.230	0.770	1.000	1.000	Total	0.140	NA
Test 2: T+ from test 1										
S+	0.413	0.850	0.150	0.351	0.062	0.923	0.100	FN	0.062	0.014
S–	0.587	0.050	0.950	0.029	0.558	0.077	0.900	FP	0.029	0.007
Totals	1.000			0.380	0.620	1.000	1.000	Total	0.091	0.021
Test 2: T– from test 1										
S+	0.006	0.850	0.150	0.006	0.001	0.100	0.001	FN	0.001	0.00075
S–	0.994	0.050	0.950	0.050	0.944	0.900	0.999	FP	0.050	0.03825
Totals	1.000			0.055	0.945	1.000	1.000	Total	0.051	0.039
Test 2										
Total										
FN										0.015
FP										0.045
Total										0.060

NA, not applicable. Misclassification categories: FN, false negative, T–| S+; FP, false positive, T+| S–; S+, positive for attribute; S–, negative for attribute; T–, test negative for attribute; T+, test positive for attribute.

intended prediction and validation sets. To have faith in model results, analysts must consider the model and the chemicals tested to determine if they are appropriately matched.

QSAR models based on the mechanism of action approach tend to rely on expert judgment to define the domain. QSAR models based on chemometric or statistical approaches tend to use similarity analysis tools where the decision is made based on formally defined similarity of chemicals in the prediction set to the chemicals in the training set. Similarity is measured as the multidimensional distance in the molecular descriptors space used as parameters of the evaluated QSAR model or by matching fragments. If the results of the similarity analysis indicate that the given QSAR model is applicable to the chemicals in the prediction set, then and only then the statistical reliability should be evaluated.

Furthermore, as described by Cronin et al. (2003a, 2003b), national and international validation centers have been established in the European Union and in the United States to validate alternative (nonanimal) methods. In this context, validation is seen as the process by which the relevance and reliability of a method for a particular purpose undergo independent assessment (Balls et al. 1995). Alternative methods include not only physicochemical and *in vitro* tests but also QSAR models and other computer-based systems for predicting toxicity. An alternative test based on physicochemical or *in vitro* data can be regarded as the combination of a test system that generates experimental data and a PM that provides an objective means of extrapolating the data to an expression of toxicity at the *in vivo* level (Worth and Balls 2001). Thus, a PM is analogous to a QSAR for an *in vivo* endpoint: the former is based on experimental physicochemical or *in vitro* data, whereas the latter is based on physicochemical descriptors. Criteria for the acceptability of PMs, which can also be applied to QSARs, are summarized below.

Acceptability Criteria

As should be evident from the discussion above, the specification of reasonable acceptability criteria for the use of QSARs in risk assessment is a multifaceted task. We try to deal with this task by grouping such criteria according to three uniting principles: *a*) basic modeling conditions, *b*) procedural steps, and *c*) reference values of performance parameters.

Basic QSAR-modeling conditions. Earlier we outlined basic modeling conditions for applicability of QSARs. Checking data for homogeneity and representativity is easily overlooked.

Homogeneity. Homogeneity means that the investigated series of compounds must have rather similar chemical and biological properties, and the mechanism of influence of X on Y must be the same. Sometimes the data set/database in question may contain many classes of compounds. These classes may be partially overlapping, barely separated, or completely resolved in the chemical descriptor (X-) space and/or biological property (Y-) space of the compounds in question. Because very strong clustering violates the assumption of homogeneity, we recommend that any QSAR modeling be commenced by studying how the compounds are clustered. PARC methods and cluster analysis techniques are ideally suited for this. For instance, a plot of the scores of the first few summary latent variables will rapidly reveal groups, trends, discontinuities, outliers, and other anomalies in the data.

Representativity. The composition of the training set and the prediction set is of crucial importance. A representative selection of compounds that well span the chemical domain of interest should be included in these sets. One way to accomplish a representative selection of compounds is through SMD (Wold et al. 1986). With this approach, test series of compounds are defined in which all major structural and chemical properties are systematically varied at the same time.

Taking into account properties of *X*- and *Y*-data. Earlier we discussed several aspects that relate to the nature and quality of the *X*- and the *Y*-data. It is of utmost importance that any knowledge about measurement noise be used in the model-building process. Any estimated "noise" in response data can be beneficially compared with the predictive power of the model. For example, if the known or expected noise is ±0.3 units for log(1/*C*), then the PRESD of *Y* should be of similar size. Also, if uncertainty estimates of many variables are available, this information can be used in the scaling of data.

Procedural steps. Before a PM/QSAR is recommended for regulatory use, it should be mandatory to carry out model validation. First, it is important to make clear what we mean by a valid model. We mean that it predicts much better than chance. In addition, it should have model coefficients that have the correct sign and with size that is proportional to their significance to the modeled process. Finally, it should be consistent with fundamental chemical, biological, and toxicologic knowledge.

To facilitate the handling of real-world data sets, a PM/QSAR should *a*) be associated with a defined endpoint that it serves to predict; *b*) take the form of an unambiguous and easily applicable algorithm for predicting a pharmacotoxicologic endpoint; *c*) ideally have a clear mechanistic basis; *d*) be accompanied by a definition of the domain of its applicability—for example, the physicochemical classes of chemicals for which it is applicable; *e*) be associated with a measure of its goodness of fit and internal goodness of prediction estimated with cross-validation or similar method to a training set of data; and *f*) be assessed in terms of its predictive power by using data that were not used in the development of the model (external validation).

In the framework of alternative tests, it is considered essential that the validation process be managed under the auspices of an organization such as the European Centre for the Validation of Alternative Methods (ECVAM) in the European Union that is independent of test method developers, who have vested interests in their own methods. Organizations such as ECVAM provide independent advice to the regulatory authorities, who have the responsibility for deciding on modifications to existing legislation (including the addition of new test methods). Additional criteria have also been developed that relate to the experimental protocols of alternative methods (Balls et al. 1995).

Reference values of performance parameters for continuous models. A third part in the compilation of acceptability criteria involves the specification of recommended values for model performance statistics such as R^2Y. The values given below must be regarded as a rule of thumb, and it might be necessary, on a case-by-case basis, to reconsider these, taking into account the purpose of the model and the variability of the underlying *X*- and *Y*-data (which place limitations on the predictive capacity). However, it is important that the model parameter criteria be defined in advance of the experimental phase of the validation study by the management team of the study. This circumvents the possibility that the criteria could be weakened, with the improper aim of "successfully" validating the method.

Proposed reference values.

- R^2Y: This limit is conditional on the Q^2Y value.
- Q^2Y: $Q^2Y > 0.5$ is generally regarded as good, and $Q^2Y > 0.9$ as excellent. These limits are highly application dependent.
- $R^2Y - Q^2Y$: This difference ought not to exceed 0.3. A substantially larger difference indicates *a*) an overfitted model (i.e., a model modeling noise); *b*) presence of irrelevant *X*-variables, or *c*) outliers in the data.
- "Background" R^2Y and Q^2Y: This consists of the intercepts of the regression lines of the response permutation testing. Results should be $R^2Y < 0.3$ and $Q^2Y < 0.05$ to indicate a valid model. These intercepts can be understood as indicating the level of "background" R^2Y and Q^2Y obtainable with random data.

Condition Number. The condition number is defined as the ratio of the largest to the smallest singular value of the *X*-matrix. When this ratio exceeds 10, this indicates that the *X*-variables are significantly correlated, and the user should refrain from using MLR. It has been recommended to use a ratio 5:1 (compounds: *X*-variables) to diminish the risk of multicollinearities among the *X*-variables.

SDEP/PRESD. The SDEP should be similar to the experimental variability of an endpoint; or, for example, if the known noise is ±0.3 units for log(1/*C*), then the SDEP (also called PRESD) of *Y* should be of similar size. Alternatively, if for instance the variability in the *Y*-data is 20% (in variance-metric), then it seems unlikely that R^2Y and Q^2Y can exceed 80% (0.8).

Reference values of performance parameters for classification models. Similar considerations apply to the Cooper statistics that are often used to assess the predictive performance of two-group classification models; that is, fixed acceptability criteria for the sensitivity, specificity, and accuracy (concordance) cannot be defined for all types of classification models because the maximal predictive performance achievable will depend on the quality of the predictor and response data. Thus, the acceptance criteria need to be established on a case-by-case basis in advance of the experimental work conducted to test the classification model.

Furthermore, the criteria should take account of the purpose of the model. For example, if the model is intended to serve as a stand-alone test, that is, a complete replacement of an animal test, then, as a minimum requirement, the Cooper statistics should be significantly greater than 50% (for a two-group model) to ensure that the model is producing predictions that are significantly better than chance (Worth and Cronin 2001b). An example is provided by an ECVAM validation study in which classification models based on *in vitro* data were assessed for their capacity to predict skin corrosion potential (Fentem et al. 1998). In this study, one of the acceptability criteria was that the sensitivity of each test be greater than 70%.

However, if a classification model is being used in a battery of tests, for example, to identify toxic chemicals that act by a certain mechanism of action, the acceptance criteria are likely to be different. For example, classification models for predicting skin corrosion potential can also be based on pH data because chemicals that are acidic or alkaline in solution are expected to be corrosive (Worth and Cronin 2001b). However, not all corrosive chemicals exert their toxic action by a pH-dependent mechanism. Thus, a model based on pH data may therefore detect a small percentage of known corrosives in a given test set (i.e., the model could have a sensitivity less than 50%), but of those chemicals it does identify as corrosive, there is a high probability of actual corrosivity [i.e., the model would have a high positive predictive value, Pr(*S*+|*P*+)]. Tests with low specificity can generate high PVP results in only two ways: *a*) the test also has very high specificity, thus generating very few false positives regardless of prevalence of *S*– items in the tested population; and *b*) a high *S*+ prevalence (thus low *S*-prevalence) will also produce few false positives because there are few negatives in the tested population. This kind of performance could be regarded as acceptable because it is understood the pH model is not a stand-alone model but is used as one component in a battery of models. Such a model is being used to identify toxic chemicals with a high degree of certainty, but it makes no predictions about nontoxic chemicals (which would need to be identified by another test in the battery).

Concluding Remarks

To increase regulatory acceptance and use of QSARs, and to enhance confidence of QSAR predictions, it is necessary to develop guidance and acceptability criteria that are not only reliable but also easy to understand and apply. This has been the intention of this article. At first sight this may seem an overly ambitious goal, particularly because many of the criteria put forward are originally statistical in nature and may therefore have a discouraging effect on the user. However, because we believe so strongly in the future use of QSARs for chemicals management, we have tried to compose a basic set of acceptability criteria so "user-friendly" in nature that each and every QSAR analyst may benefit from them.

In summary, we emphasize the value of predictive QSAR models in future chemicals management, including priority setting, risk assessment, and classification and labeling. These models can be seen as simplifications (approximations) of complicated functional relationships that often prevail between chemical and biological properties of compounds. Provided that QSARs are applied with care and common sense and are developed by fulfilling the basic acceptability criteria outlined-here, they constitute an important and powerful tool definitely deserving a slot in the risk assessor's toolbox.

References

Albano C, Dunn WG III, Edlund U, Johansson E, Nordén B, Sjöström M, et al. 1978. Four levels of pattern recognition. Anal Chem Acta 103:429–443.

Andersson PM, Sjöström M, Wold S, Lundstedt T. 2000. Comparison between physicochemical and calculated molecular descriptors. J Chemomr 14:629–642.

[Anonymous]. 1995. QSAR for prediction of fate and effects of chemicals in the environment. Environmental Technologies RTD Programme (DGXII/D-1). Contract EV5V-CT92-0211. Brussels:Commission of the European Union.

———. 1999. Fate and activity modeling of environmental pollutants using structure-activity relationships (FAME). Contract number ENV4-CT96-0221. Brussels:Environment and Climate Programme of the European Union.

Atkinson AC. 1985. Plots, Transformations and Regression. Oxford:Clarendon Press.

Balls M, Blaauboer BJ, Fentem JH, Bruner L, Combes RD, Ekwall B, et al. 1995. Practical aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM workshop 5. Altern Lab Anim 23:129–147.

Baroni M, Costatino G, Cruciani G, Riganelli D, Valigi R, Clementi S. 1993. Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. Quant Struct-Act Rel 12:9–20.

Berglund A, De Rosa MC, Wold S. 1997. Alignment of flexible molecules at their receptor site using 3D descriptors and Hi-PCA. J Comput Aid Mol Des 11:601–612.

Blum DJW, Speece RE. 1990. Determining chemical toxicity to aquatic species. Environ Sci Technol 24:284–293.

Box GEP, Hunter WG, Hunter JS. 1978. Statistics for Experimenter. New York:Wiley.

Burden FR, Rosewarne BR, Winkler DA. 1997. Predicting maximum bioactivity by effective inversion of neural networks using genetic algorithms. Chemomr Intell Lab 38:127–137.

Burnham AJ, MacGregor JF, Viveros R. 1999. A statistical framework for multivariate latent variable regression methods based on maximum likelihood. J Chemomr 13: 49–65.

———. 2001. Interpretation of regression coefficients under a latent variable regression model. J Chemomr 15:265–284.

Burnham AJ, Viveros R, MacGregor JF. 1996. Frameworks for latent variable multivariate regression. J Chemomr 10:31–45.

Cooper JA, Saracci R, Cole P. 1979. Describing the validity of carcinogen screening tests. Br J Cancer 39:87–89.

Cronin MTD, Jaworska J, Walker JD, Comber M, Watts CD, Worth AP. 2003b. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ Health Perspect 111:1391–1401.

Cronin MTD, Schultz TW. 2003. Pitfalls in QSAR. J Mol Struct 633:39–51.

Cronin MTD, Sinks GD, Schultz TW. 2000. Modeling of toxicity to the ciliate tetrahymena pyriformis: the aliphatic carbonyl domain. In: Forecasting the Environmental Fate and Effects of Chemicals (Rainbow PS, Hopkins SP, Crane M, eds). Chichester, UK:Wiley, 113–124.

Cronin MTD, Walker JD, Jaworska J, Comber M, Watts CD, Worth AP. 2003a. Use of QSAR relationships in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect 111:1376–1390.

Cruciani G, Watson KA. 1994. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. J Med Chem 37:2589–2601.

Davis L. 1991. Handbook of Genetic Algorithms. New York:Van Nostrand Reinhold.

Deneer JW, Sinnige TL, Seinen W, Hermens JLM. 1987. Quantitative structure-activity relationships for the toxicity and bioconcentration factor of nitrobenzene derivatives towards the guppy. Aquat Toxicol 10:115–129.

Deneer JW, van Leeuwen CJ, Seinen W, Maas-Diepeveen JL, Hermens JLM. 1989. QSAR study of the toxicity to nitrobenzene derivatives towards *Daphnia magna, Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. Aquat Toxicol 15:83–97.

Denham MC. 1997. Prediction intervals in partial least squares. J Chemomr 11:39–52.

Devillers J. 1996. Genetic algorithms in computer-aided molecular design. In: Genetic Algorithms in Molecular Modeling (Devillers J, ed). London:Academic Press, 131–157.

Devillers J, Balaban AT. 1999. Topological Indices and Related Descriptors in QSAR and QSPR. Amsterdam:Gordon Breach Scientific Publishers, 811.

Draper NR, Smith H. 1981. Applied Regression Analysis. New York:Wiley and Sons.

Dunn WJ III. 1989. Quantitative structural-activity relationships. Chemomr Intell Lab 6:181–189.

Efron B. 1982. The Jackknife, the Bootstrap and Other Resampling Planes. Philadelphia:Society for Industrial and Applied Mathematics.

Efron B, Gong G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37:36–48.

Efron B, Tibshirani RJ. 1993. An Introduction to the Bootstrap. London:Chapman & Hall.

Eriksson L, Hermens JLM, Johansson E, Verhaar HJM, Wold S. 1995. Multivariate analysis of aquatic toxicity data with PLS. Aquat Sci 57:217–241.

Eriksson L, Johansson E. 1996. Multivariate design and modeling in QSAR. Chemomr Intell Lab 34:1–19.

Eriksson L, Johansson E, Kettaneh-Wold N, Wikström C, Wold S. 2000b. Design of Experiments—Principles and Applications. Umea, Sweden:Umetrics AB.

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. 2001. Multi-and Megavariate Data Analysis—Principles and Applications. Umea, Sweden:Umetrics AB.

Eriksson L, Johansson E, Lindgren F, Sjöström M, Wold S. 2002. Megavariate analysis of hierarchical QSAR data. J Comp Aid Mol Des 16:711–726.

Eriksson L, Johansson E, Lindgren F, Wold S. 2000. GIFI-PLS: modeling of non-linearities and discontinuities in QSAR. Quant Struct-Act Rel 19:345–355.

Eriksson L, Johansson E, Müller M, Wold S. 2000a. On the selection of training set in environmental QSAR when compounds are clustered. J Chemomr 14:599–616.

Eriksson L, Johansson E, Wold S. 1997. QSAR model validation. In: Quantitative Structure-Activity Relationships in Environmental Sciences—VII (Chen F, Schüürmann G, eds). Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, 24–28 June 1996, Elsinore, Denmark. Pensacola, FL:SETAC Press, 381–397.

Feinstein AR. 1975. Clinical biostatistics XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests. Clin Pharmacol Ther 17:104–116.

Fentem JH, Archer GEB, Balls M, Botham PA, Curren RD, Earl LK, et al. 1998. The ECVAM international validation study on *in vitro* tests for skin corrosivity. 2. Results and evaluation by the management team. Toxicol In Vitro 12:483–524.

Giraud E, Luttman C, Lavelle F, Riou JF, Mailliet P, Laoui A. 2000. Multivariate data analysis using D-optimal designs, PLS and RSM. A directional approach for the analysis of farnesyltransferase inhibitors. J Med Chem 43:1807–1816.

Golbraikh A, Tropsha A. 2002. Beware of q2! J Mol Graph Model 20:269–276.

Goldberg D.E. 1989. Genetic Algorithms in Search, Optimization & Machine Learning. New York:Addison-Wesley.

Gombar V. 1996. U.S. Patent 6 036 349. Method and apparatus for validation of model-based predictions.

Gramatica P, Consonni V, Todeschini R. 1999. QSAR study of the tropospheric degradation of organic compounds. Chemosphere 38:1371–1378.

Gramatica P, Corradi M, Consonni V. 2000. Modeling and prediction of soil sorption coefficients of non-ionic organic pesticides by different sets of molecular descriptors. Chemosphere 41:763–777.

Gramatica P, Navas N, Todeschini R. 1998. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). Chemomr Intell Lab 40:53–63.

Gramatica P, Papa E. 2003. QSAR modeling of bioconcentration factor by theoretical molecular descriptors. Quant Struct-Act Rel 22:374–385.

Hibbert DB. 1993. Genetic algorithms in chemistry. Chemomr Intell Lab 19:277–293.

Hong H, Tong W, Fang H, Shi L, Qian X, Wu J, et al. 2002. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated computational approach. Environ Health Perspect 110:29–36.

Höskuldsson A. 1996. Prediction Methods in Science and Technology. Copenhagen:Thor Publishing.

Jackson JE. 1991. A User's Guide to Principal Components. New York:John Wiley.

Jongman RGH, ter Braak CJF, van Tongeren OFR. 1987. Data Analysis in Community and Landscape Ecology. Wageningen, the Netherlands:Pudoc.

Jouan-Rimbaud D, Massart DL, de Noord OE. 1996. Random correlation in variable selection for multivariate calibration with a genetic algorithm. Chemomr Intell Lab 35:213–220.

Jurs PC. 2002. ADAPT—Automated Data Analysis and Pattern Recognition Toolkit. University Park, PA:Pennsylvania State University. Available: http://research.chem.psu.edu/pcjgroup/ADAPT.html [accessed 23 April 2002].

Karelson M. 2000. Molecular Descriptors in QSAR/QSPR. New York:Wiley-InterScience.

Katritzky AR, Lobanov VS, Karelson M. 1994. CODESSA, Reference Manual. Gainesville, FLUniversity of Florida. Available: http://www.semichem.com/codessarefs.html [accessed 19 April 2002].

Kubinyi H. 1994a. Variable selection in QSAR studies. I. An evolutionary algorithm. Quant Struct-Act Rel 13:285–294.

———. 1994b. Variable selection in QSAR Studies. II. A highly efficient combination of systematic search and evolution. Quant Struct-Act Rel 13:393–401.

———. 1996. Evolutionary variable selection in regression and PLS analyses. J Chemomr 10:119–133.

Kulkarni A, Hopfinger AJ, Osborne R, Bruner LH, Thompson ED. 2001. Prediction of eye irritation of organic chemicals using membrane-interaction QSAR analysis. Toxicol Sci 59:335–345.

Langer T. 1994. Molecular similarity determination of heteroaromatics using CoMFA and multivariate data analysis. Quant Struct-Act Rel 13:402–405.

Leardi R. 1994. Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. J Chemomr 8:65–79.

Leardi R, Boggia R, Terrile M. 1992. Genetic algorithms as a strategy for feature selection. J Chemomr 6:267–281.

Lindgren F. 1994. Third Generation PLS—Some Elements and Applications [PhD Thesis]. Umeå, Sweden:Umeå University.

Linusson A, Gottfries J, Lindgren F, Wold S. 2000. Statistical molecular design of building blocks for combinatorial chemistry. J Med Chem 43:1320–1328.

Livingstone DJ. 2000. The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 40:195–209.

Marengo E, Todeschini, R. 1992. A new algorithm for optimal, distance—based experimental design. Chemomr Intell Lab 16:37–44.

Martens H, Martens M. 2000. Modified jack-knife estimation of parameter uncertainty in bilinear modeling (PLSR). Food Qual Prefer 11:5–16.

Martin YC, Lin CT, Hetti C, DeLazzer J. 1995. PLS analysis of distance matrices to detect nonlinear relationships between biological potency and molecular properties. J Med Chem 38:3009–3015.

McDowell RM, Jaworska J. 2002. Bayesian analysis and inference of QSAR predictive model results. SAR QSAR Environ Res 13:111–125.

Mekenyan O, Bonchev D. 1986. OASIS method for predicting biological activity of chemical compounds. Acta Pharm Jugosl 36:225–237.

Mullet GM. 1976. Why regression coefficients have the wrong sign. J Qual Technol 8:121–126.

Nendza M, Müller M. 2000. Discriminating toxicant classes by mode of action: 2. Physico-chemical descriptors. Quant Struct-Act Rel 19:581–598.

OECD. 1998. Harmonized Integrated Hazard Classification System for Human Health and Environmental Effects of Chemical Substances. Paris:Organisation for Economic Cooperation and Development.

Pet-Edwards J, Haimes Y, Chankong V, Rosenkranz H, Ennever F. 1989. Risk Assessment and Decision Making Using Tests Results—The Carcinogenicity Prediction and Battery Selection Approach. New York:Plenum Press, 211.

Rogers D, Hopfinger AJ. 1994. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 34:854–866.

Shao J. 1993. Linear model selection by cross-validation. J Am Stat Assoc 88:486–494.

Shi LM, Tong W, Fang H, Perkins R, Wu J, Tu M, et al. 2002. An integrated "four-phase" approach for priority setting of endocrine disruptors—phase I and II for prediction of potential estrogenic endocrine disruptor. SAR/QSAR Environ Res 13(1):69–88.

Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. 1998. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemomr Intell Lab 44: 229–244.

Sjöström M, Lindgren Å, Uppgård L. 1997. Joint multivariate quantitative structure-property and structure-activity relationships for a series of technical nonionic surfactants. In: Quantitative Structure-Activity Relationships in Environmental Sciences—VII (Schüürmann G, Chen F, eds). Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, 24–28 June 1996, Elsinore, Denmark. Pensacola, FL:SETAC Press, 435–449.

Stuper AJ, Jurs PC. 1976. ADAPT: A computer system for automated data analysis using pattern recognition techniques. J Chem Inf Comput Sci 16:99–105.

Todeschini R. 1997. MOBY DIGS—Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 2.1 for Windows. Milan:Talete Srl.

Todeschini R, Consonni V. 2000. Handbook of Molecular Descriptors. Weinheim:Wiley-VCH.

Todeschini R, Consonni V, Pavan M. 2001. DRAGON—Software for the Calculation of Molecular Descriptors. Release 1.12 for Windows. Available: http://www.disat.unimib/chm [accessed 25 March 2002].

Todeschini R, Gramatica P. 1997. 3D-modelling and prediction by WHIM descriptors. Part 6. Applications of WHIM descriptors in QSAR studies. Quant Struct-Act Rel 16:120–125.

Todeschini R, Maiocchi A, Consonni V. 1999. The K correlation index: theory development and its application in chemometrics. Chemomr Intell Lab 46:13–29.

Tong W, Perkins R, Fang H, Hong H, Xie Q, Branham W, et al. 2002. Development of quantitative structure-activity relationships (QSARs) and their use for priority setting in testing strategy of endocrine disruptors. Regul Res Perspect 1(3):1–16.

Topliss JG, Edwards RP. 1979. Chance factors in studies of quantitative structure-activity relationships. J Med Chem 22:1238–1244.

Tosato ML, Piazza R, Chiorboli C, Passerini L, Pino A, Cruciani G, et al. 1992. Application of chemometrics to the screening of hazardous chemicals. Chemomr Intell Lab 16:155–167.

Tropsha A, Gramatica P, Gombar VJ. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. Quant Struct-Act Rel 22:69–77.

Tysklind M, Andersson P, Haglund P, Van Bavel B, Rappe C. 1995. Selection of polychlorinated biphenyls for use in quantitative structure-activity relationships. SAR QSAR Environ Res 4:11–19.

Van der Voet H. 1994. Comparing the predictive accuracy of models using a simple randomization test. Chemomr Intell Lab 25:313–323.

Vankeerberghen P, Smeyers-Verbeke J, Leardi R, Karr CL, Massart DL. 1995. Robust regression and outlier detection for non-linear models using genetic algorithms. Chemomr Intell Lab 28:73–87.

Verhaar HJM, Eriksson L, Sjöström M, Schüürmann G, Seinen W, Hermens JLM. 1994. Modeling the toxicity of organophosphates: a comparison of the multiple linear regression and PLS regression methods. Quant Struct-Act Rel 13:133–143.

Wahlström B. 1988. The need for new strategies—the OECD existing chemicals program. In: The Use of QSAR for Chemicals Screening—Limitations and Possibilities. Report 8/1988. Stockholm:National Chemicals Inspectorate, 1–17.

Wehrens R, Buydens, LMC. 1998. Evolutionary optimization: a tutorial. Trends Analyt Chem 17:193–203.

Wehrens R, Putter H, Buydens LMC. 2000. The bootstrap: a tutorial. Chemomr Intell Lab 54:35–52.

Wold H. 1982. Soft modeling. The basic design and some extensions. In: Systems under Indirect Observation, Vols I, II (Jöreskog KG, Wold H, eds). Amsterdam:North-Holland, 18–34.

Wold S. 1992. Nonlinear partial least squares modeling. II. Spline inner relation. Chemomr Intell Lab 14:71–84.

Wold S, Albano C, Dunn WJ III, Esbensen K, Hellberg S, Johansson E, et al. 1983. Pattern recognition: finding and using regularities in multivariate data. In: Food Research and Data Analysis (Martens H, Russwurm H, eds). Essex, UK: Applied Science Publishers, 147–188.

Wold S, Dunn WJ III. 1983. Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. J Chem Inf Comput Sci 23:6–13.

Wold S, Johansson E, Cocchi M. 1993. PLS—partial least squares projections to latent structures. In: 3D-QSAR in Drug Design, Theory, Methods, and Applications (Kubinyi H, ed). Leiden:ESCOM Science Publishers, 523–550.

Wold S, Josefson M. 2000. Multivariate calibration of analytical data. In: Encyclopedia of Analytical Chemistry (Meyers RA, ed). Chichester, UK:John Wiley & Sons, 9710–9736.

Wold S, Sjöström M, Carlson R, Lundstedt T, Hellberg S, Skagerberg B, et al. 1986. Multivariate design. Anal Chem Acta 191:17–32.

Worth AP, Balls M. 2001. The importance of the prediction model in the development and validation of alternative tests. Altern Lab Anim 29:135–143.

Worth AP, Cronin MTD. 2000. Embedded cluster modeling: a novel QSAR method for generating elliptic models of biological activity. In: Progress in the Reduction, Refinement and Replacement of Animal Experimentation (Balls M, van Zeller A-M, Halder ME, eds). Amsterdam:Elsevier Science, 479–491.

———. 2001a. The use of bootstrap resampling to assess the uncertainty of Cooper statistics. Altern Lab Anim 29: 447–459.

———. 2001b. The use of pH measurements to predict the potential of chemicals to cause acute dermal and ocular toxicity. Toxicology 169:119–131.

———. In press. The use of discriminant analysis, logistic regression and classification tree analysis in the devlopment of classification models for human health effects. J Mol Struct.

Xu L, Zhang WJ. 2001. Comparison of different methods for variable selection. Anal Chim Acta 446:477–483.