JOURNAL OF CHEMICAL INFORMATION AND MODELING

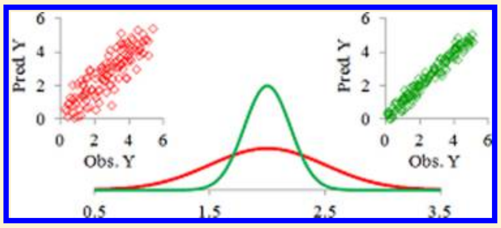
Article

pubs.acs.org/jcim

How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models

Mark C. Wenlock*[,†] and Lars A. Carlsson[‡]

[†]Drug Safety & Metabolism, AstraZeneca R&D Alderley Park, Macclesfield, Cheshire, SK10 4TF, U.K.
[‡]Drug Safety & Metabolism, AstraZeneca R&D Mölndal, 431 83, Mölndal, Sweden



ABSTRACT: We consider the impact of gross, systematic, and random experimental errors in relation to their impact on the predictive ability of QSAR/QSPR DMPK models used within early drug discovery. Models whose training sets contain fewer but repeatedly measured data points, with a defined threshold for the random error, resulted in prediction improvements ranging from 3.3% to 23.0% for an external test set, compared to models built from training sets in which the molecules were defined by single measurements. Similarly, models built on data with low experimental uncertainty, compared to those built on data with higher experimental uncertainty, gave prediction improvements ranging from 3.3% to 27.5%.

INTRODUCTION

From the perspective of drug discovery, quantitative structure activity (or property) relationship (QSA(P)R) models are sought that mathematically relate a set of predictor variables (typically electronic, fragmental, spatial, steric, structural, thermodynamic, and topological based) for a molecule to an experimentally determined response variable. These models can then be used to quantitatively predict the response variable for the next set of molecules being designed or to triage molecules for subsequent screening. Within the drug discovery process, it is common for empirically based progression criteria to exist for a particular response variable. This allows QSA(P)R predictions to be used to inform the decision regarding the risks associated with the synthesis or screening progression for a particular molecule. Although these decisions can be multifarious, the application of QSA(P)R models is intended to minimize the efforts spent on molecules that may not subsequently pass the progression criteria.

In practice, not all QSA(P)R models are useful for decision-making purposes and this can be for several reasons,[1−4] but there are three practical aspects that should be highlighted. First, it is assumed that molecules that are similar in the predictor variable space of the molecules within the model's training set will have similar activities (or properties). For experimental variables (e.g., aqueous solubility) that are predominantly influenced by bulk molecular properties, this assumption may be reasonable. However, when the model's predictor variable space is unable to account for subtle molecule−protein interactions associated with certain experimental variables, this assumption becomes less reliable. Second, each modeling algorithm is able to mathematically relate, to different extents, the predictor variables to the response variables. The third aspect presumes that the experimentally determined response variable that is being modeled is the true value.[5] The true value will be given by the population mean value of all possible measurements (which also assumes that there are no systematic errors in the measurements). As no physical quantity can be measured with absolute certainty, it is unlikely that the value of a single measurement will be equal to the true value for a molecule.

Within AstraZeneca, automated procedures will, at a predefined time interval, take any new response variable data and automatically rebuild the QSA(P)R model.[6,7] Thereby the QSA(P)R models are kept up to date by being trained on the most current molecules, as well as being relevant for legacy molecules. The modeling algorithms use the training set response variable data, which may be a single measurement or the mean of a sample of measurements, and attempt to build a model. Thus, it is not the true values of the response variable for the training set of molecules that the modeling algorithms are trying to relate the predictor variables to, rather it is the estimate of the true value that has inherent uncertainty. It follows that the model's prediction of the external test set will have uncertainty equal to or greater than that contained within the training set. When comparing these predictions against the observed data for an external test set, the predictive ability of the model will be a function of the propagated errors associated with the predictions and also the observed measurements. Therefore, a model's predictive ability needs to be contextualized against the experimental error in the determination of the response variable. Although beyond the scope of this work, it should be noted that advanced machine learning techniques like Gaussian Process methods explicitly estimate the error in the experimental data and hence can estimate the prediction error from the true value.[8] Qualitatively, the predictive ability of a model will be low if the experimental error is high and may be high if the experimental error is low.

This study investigates the impact of experimental errors on the predictive ability of QSA(P)R models, built on AstraZeneca

Received: September 4, 2014
Published: November 18, 2014

data sets, for eight drug metabolism and pharmacokinetic (DMPK) response variables used within early drug discovery. These include: (i) human hepatocyte intrinsic clearance (human hep CL_{int}), (ii) human microsome intrinsic clearance (human mic CL_{int}), (iii) the extent of human plasma protein binding (human PPB), (iv) distribution coefficients between octan-1-ol/pH7.4 aqueous buffer ($\log_{(10)} D_{7.4}$), (v) the extent of rat plasma protein binding (rat PPB), (vi) rat hepatocyte intrinsic clearance (rat hep CL_{int}), (vii) aqueous (pH7.4) solubility from an initial dried DMSO form (solubility (dried DMSO)), and (viii) aqueous (pH7.4) solubility from an initial solid form (solubility (solid)). The impact of assay screening strategies, where the aim has been to characterize such DMPK response variables on all compounds synthesized within early drug discovery, has come at the expense of having poorer estimates of the true values, as often only single measurements have been made. This experimental estimate may be sufficient for a progression decision to be made on a particular compound, but the collective effect to QSA(P)R modeling is to increase the uncertainty in the training sets being used, irrespective of the methodology being employed. Therefore, it is important to quantify the effect of such strategies on the ability of QSA(P)R models to subsequently influence the drug discovery process.

■ MATERIALS AND METHODS

Each of the eight DMPK response variables has been experimentally determined across multiple AstraZeneca research sites. Some of the experimental data (e.g., log $D_{7.4}$ data) had been generated over nearly three decades using procedures typically employed within the industry during this time. Therefore, the data sets for each response variable contain measurements pooled from different assays that are considered, for the purpose of QSA(P)R modeling, to be comparable. The human hep CL_{int} data set contains experimental data from three assays where incubations were run at 37 °C for up to 120 min, using either cryopreserved or fresh hepatocytes, with values ranging from 3.00 to 300.00 μL/min/1 × 10^6 cells. The experimental procedures are similar to those described by Temesi et al.[9] The human hep CL_{int} is a first-order metabolic degradation rate constant, determined at low substrate concentrations, divided by the human hepatocyte concentration used within the incubation. It provides an estimate of the ratio of the maximum enzymatic rate (i.e., V_{m}) to the Michaelis–Menten constant (i.e., K_{m}) for the molecule. The same applies to the human mic CL_{int} and rat hep CL_{int}, albeit the former uses human microsomal protein and the latter rat hepatocytes. The human mic CL_{int} data set contains experimental data from five assays where incubations were run at 37 °C for up to 60 min, using cryopreserved human liver microsomal protein, with values ranging from 3.00 to 300.00 μL/min/mg of protein. The experimental procedures are similar to those exemplified by Temesi et al.[9] for hepatocytes, except for the use of 1 mg/mL microsomal protein rather than 1 × 10^6 cells. The human PPB data set contains experimental data from 11 equilibrium dialysis-based assays run at 37 °C for greater than 4 h, using pooled neat plasma, with values ranging from 10.00% to 99.95% bound. The experimental procedures are similar to those reported by Fessey et al.[10] The log $D_{7.4}$ data set contains experimental data from nine assays run at room temperature (ranging from 20 to 25 °C), measuring shake-flask octan-1-ol/aqueous pH7.4 (0.01 or 0.02 M) phosphate buffer biphasic distribution coefficients, using either solid or dried DMSO starting material, with values ranging from −1.50 to 4.50. The experimental procedures are similar to those described by Wenlock et al.[11] The rat hep CL_{int} data set contains experimental data from six assays where incubations were run at 37 °C for up to 120 min, using either cryopreserved or fresh hepatocytes, with values ranging from 3.00 to 300.00 μL/min/1 × 10^6 cells. The experimental procedures are similar to those exemplified by Temesi et al.[9] The rat PPB data set contains experimental data from 11 equilibrium dialysis-based assays run at 37 °C for greater than 4 h, using pooled neat plasma from various strains, with values ranging from 10.00% to 99.95% bound. The experimental procedures are similar to those reported by Fessey et al.[10] The solubility (dried DMSO) data set contains data from seven shake-flask pH7.4 phosphate buffer assays, run at room temperature (ranging from 22 to 25 °C), using dried DMSO starting material with a centrifugation separation technique, with values ranging from 0.10 to 1500.00 μM. The experimental procedures are similar to those described by Alelyunas et al.[12] The solubility (solid) data set contains data from five shake-flask pH7.4 phosphate buffer assays, run at room temperature (ranging from 22 to 25 °C), using solid starting material with a centrifugation separation technique, with values ranging from 0.10 to 1500.00 μM. The experimental procedure is similar to those exemplified by Wenlock et al.[13]

Pipeline Pilot[14] scripts were written for data preparation, analysis, and modeling. Any measurement that was less than or greater than the specified experimental dynamic range was excluded. In addition, molecules with repeat measurements whose standard deviation (stdev) was greater than twice the typical stdev for that response variable were also removed. Where molecules had repeat, acceptable measurements the mean value was considered. The number of molecules remaining after these steps, along with the percentage of molecules defined by a single measurement, is shown in Table 1. The variation in these percentages is indicative of different screening strategies used over time at AstraZeneca. The statistical package R (accessed via Pipeline Pilot) was used to perform all principal component analyses and random forest recursive partitioning was used to build the QSA(P)R models.[15] This random forest methodology had a practical limit that restricted the size of the model training sets to 35 000 molecules. Where a training set size exceeded this limit, 35 000

Table 1. Overview of Data Sets for the Eight DMPK Response Variables

response variable	number of molecules	number of results	number of molecules to consider	percentage of data set with a single measurement
human hep CL_{int}	10668	22588	9819	40
human mic CL_{int}	32492	47566	31215	74
human PPB	61356	80725	59852	89
log $D_{7.4}$	115441	140662	113339	93
rat hep CL_{int}	39112	55969	36807	77
rat PPB	16476	23738	16037	85
solubility (dried DMSO)	44256	49043	42821	95
solubility (solid)	38722	42736	36256	95

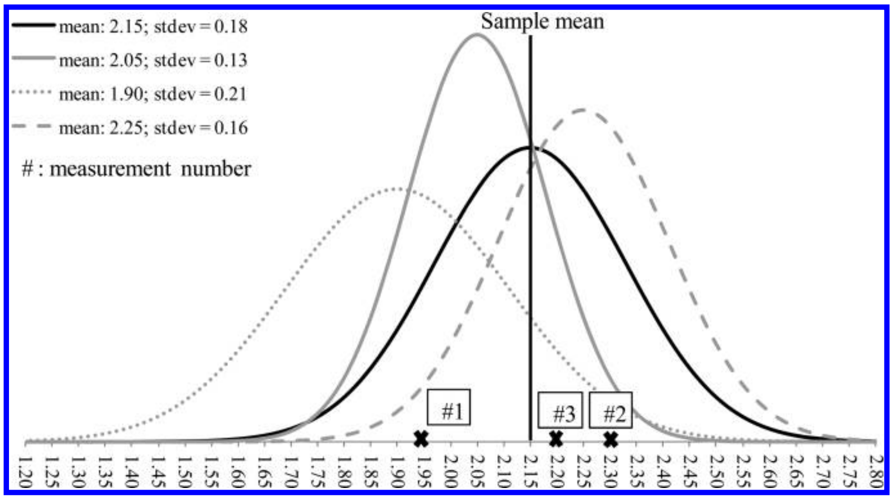


Figure 1. Sample normal distribution and three possible population normal distributions.

molecules were randomly chosen. Molecular data sets that were split into training sets and external test sets used a randomly chosen minimum 80/20 split. In all instances, any random selection of training and test data sets was repeated prior to any new model being built. Predictor variables were made up of topological, geometrical, electronic, and functional group descriptors from the AstraZeneca and Dragon descriptor sets.[16,17] The response variables were modeled on a logarithmic base 10 scale; for human and rat PPB the percentage bound data was converted to a $\log_{10}$(% bound/% free) scale.[10] Various high-quality (HQ) data and low-quality (LQ) data definitions have been explored. In all cases the HQ data and LQ data training sets were used to build a random forest recursive partitioning model via Pipeline Pilot. The predictive ability of these QSA(P)R models was quantified by the magnitude of the root mean squared error in prediction (RMSEP) for a HQ data external test set of molecules.

■ RESULTS AND DISCUSSIONS

True Value versus Sample Mean. For a particular molecule, consider the three repeat measurements 1.95, 2.30, and 2.20 for a particular response variable. Assuming a normal distribution for these repeats, the mean of this sample can be estimated to be 2.15 and the stdev will be 0.18; the normal distribution curve for such a mean and stdev is shown by the black curve in Figure 1. The true value will be given by the population mean value of all possible measurements, assuming the absence of systematic errors in the measurements. This sample of measurements can be used to estimate the range, which includes the true value; for this small sample the 95% confidence range is 2.15 ± 0.45, as determined by eq 1.[5]

$$\bar{x} \pm \left(\frac{t_{n-1} \times \text{stdev}}{\sqrt{n}}\right) \qquad (1)$$

Where $\bar{x}$ is the sample mean, n is the number of repeat measurements, and the t value, which is dependent on the number of degrees of freedom (i.e., $n - 1$), is 4.30 when considering the 95% confidence intervals. For illustration purposes, if all possible measurements could be made it is feasible that the distribution could have a mean of 2.05 and a stdev of 0.13 (solid gray curve in Figure 1), a mean of 1.90 and a stdev of 0.21 (small dashed gray curve in Figure 1), or a mean of 2.25 and a stdev of 0.16 (dashed gray curve in Figure 1). If the sample mean is used as the true value for a particular response variable for a particular molecule, then in these three scenarios it would be out by 0.10, 0.25, and −0.10, respectively. In practice, the true value is unknown but it is likely that the mean value of a sample of repeat measurements is a reasonable estimate, which will be better than just considering a single measurement. Intuitively, the larger the sample of repeat measurements performed, the more reliable the estimate of the molecule's true value. Within drug discovery there will exist practical resource and time constraints that tend to limit the number of repeat measurements performed on each molecule. It is very common within AstraZeneca to find that the majority of the data sets for response variables are being defined with a single measurement (see Table 1), and it is important to appreciate the effect of this on QSA(P)R models.

Definition of a Good QSA(P)R Model. The quality of a QSA(P)R model can be determined by how successful it is at facilitating the decisions regarding the synthesis or screening progression of a particular molecule. With respect to the correlation between the predicted values and the observed response variable values for an external test set, it is important to define the maximum RSMEP that will distinguish a good model from a moderate or poor model. It is proposed that this RMSEP criterion should be less than or equal to 0.30. Figure 2 illustrates such a plot, where the correlation between the predicted response variable for an external test set (y-axis) and the actual experimental values (x-axis) has an RMSEP of ~0.30. On inspection it can be seen that across the dynamic range, the

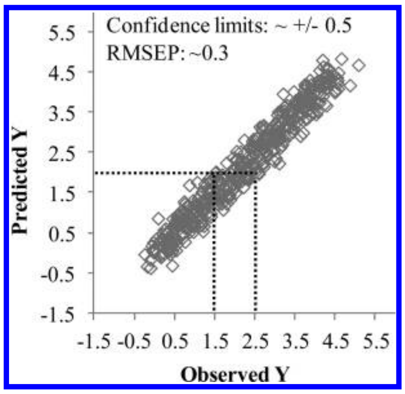


Figure 2. Hypothetical observed versus predicted plot.

maximum spread in the agreement between the x and y data is ~1.00 unit. In other words, the predicted Y estimates the observed Y within ± ~0.50 units. The ability of a QSA(P)R model to predict the experimentally determined response variable for new molecules to within ±0.50 units, on greater than or equal to 99% of the occasions, is considered to be good and would correspond to a RMSEP of ~0.30.

The plot in Figure 2 is a hypothetical simulation of randomly chosen x and y values from 500 identical normally distributed x and y populations, with each population stdev equal to 0.20. Consider the population that has a mean of 2.00, then the randomly chosen value is likely to fall between 1.80 and 2.20 68% of the time and between 1.60 and 2.40 95% of the time (note that if the stdev is equal to 0 then the correlation in Figure 2 would be ideal). This simulation highlights that the RMSEP is approximately equal to the squared root of the error propagation of the population variance in the predicted and experimentally determined response variables (i.e., $(0.20^2 + 0.20^2)^{0.5} = 0.28$).

The x-axis of the plot in Figure 2 can be thought of as representing the experimental response variables for an external test set of molecules, where the uncertainty associated with each value is equivalent to a constant stdev of 0.20. The y-axis can be considered as being the QSA(P)R model predictions for the response variable values for the external test set. It assumes that the predictor variables and modeling algorithms perfectly modeled the experimental response variables in the training set, but the uncertainty associated with each experimental value also equated to a constant stdev of 0.20. Therefore, each prediction of a molecule's true value for the response variable is associated with an error of 0.20. In practice, there is a high likelihood that modeling errors will also be present that will propagate through into a worsening of the RMSEP. Reflecting on this definition of what makes a good QSA(P)R model, the likelihood of building good models is low if the experimental uncertainty in a response variable measurement is typically greater than a stdev of 0.20.

An appreciation of a response variable's gross, systematic and random experimental errors is critical in managing the expectations associated with a potential QSA(P)R model. With respect to repeat measurements, random errors appear as measurements tending to fall randomly either side of an average value. Systematic errors appear as measurements always being biased either negatively or positively. Gross errors are associated with a serious deviation from the validated experimental procedures (i.e., an instrument malfunction) appearing as widely varying measurements, more so than would be expected from just random variability. From a QSA(P)R perspective, it is considered that gross and systematic experimental errors should be removed from modeling data sets. For the eight DMPK response variables being considered, they tend to be medium, partially automated throughput assays (i.e., hundreds of compounds per week) and, on occasion, gross errors can appear within the data sets that are subsequently used for modeling.[6] These response variables have been important for many years and, over time, systematic errors can appear in the data sets due to differences within the experimental procedures.

Gross Experimental Errors. Table 1 indicates that the number of experimental data points being considered amounts to over 463 000, collected over many years. It can be assumed that each data point would have been generated against a validated experimental procedure and only reported if acceptance criteria had been met. Repeat measurements may have been generated in different assays. It is not implicit that the measurement is good just because the acceptance criteria have been met; all that is implied is that to the best knowledge of the experimentalist there were no obvious reasons not to accept the data and that there was a high likelihood that the measurement was good. However, it is prudent to assume that the data sets may contain some gross errors; although for the purposes of this study it is impractical to consider the accuracy of every single measurement. For those molecules with response variable values defined by a single measurement, this value has been considered to be good provided that it falls within the assay's dynamic ranges. It is only for molecules with repeat measurements that potential gross experimental errors can be identified. Pragmatically, these could be identified where the sample stdev is greater than, or equal to, a certain fold over the typical stdev seen for the response variable for all available molecules, with three or more repeat measurements. For the purpose of this study a twofold criterion has been assumed for all data sets. This gross experimental error criterion could be considered to be subjective because, by chance, some molecules will have repeat measurements that exceed it. However, experimentalists are unlikely to knowingly report repeat measurements that differ widely and such instances most likely occur due to human oversight. The typical stdevs are estimated from the linear regression line fitted to the mean and stdev values for molecules with three or more repeat measurements for a particular response variable. Figure 3 shows such plots for the (i) human PPB (i.e., $\log_{10}$(% bound/% free)), (ii) log $D_{7.4}$, (iii) solubility (dried DMSO) (i.e., $\log_{10}$(solubility (dried DMSO)), and (iv) human mic CL_{int} (i.e., $\log_{10}$(human mic CL_{int})) response variables. Table 2 summarizes the details for all eight DMPK response variables including the regression line coefficients.

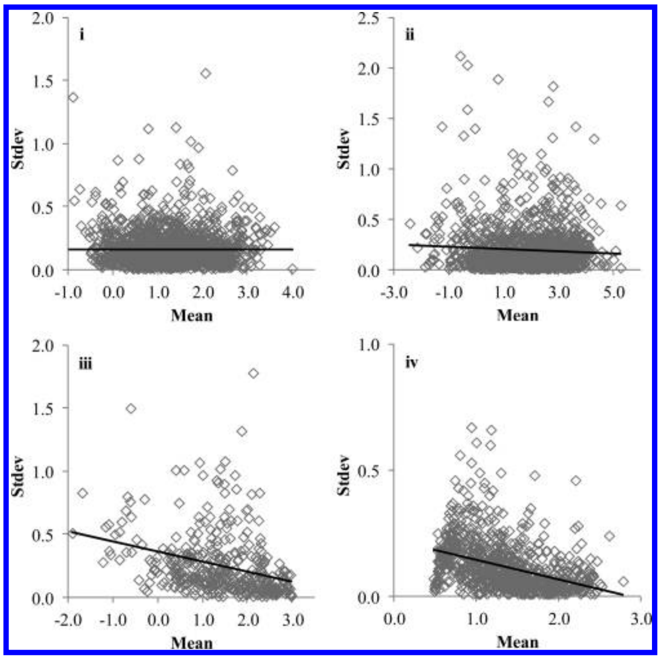


Figure 3. Scatter plots of sample mean and stdev for molecules with three or more repeat measurements for (i) human PPB (i.e., $\log_{10}$(% bound/% free)), (ii) log $D_{7.4}$, (iii) solubility (dried DMSO) (i.e., $\log_{10}$(solubility (dried DMSO)), and (iv) human mic CL_{int} (i.e., $\log_{10}$(human mic CL_{int})).

Table 2. Response Variable Summaries of the Linear Regression Fit of the Sample Mean and Stdev for Molecules with Three or More Repeat Measurements

response variable	number of molecules with three or more repeat measurements	slope	intercept
human hep CL_{int}	571	-0.02^a	0.13
human mic CL_{int}	841	-0.08^a	0.22
human PPB	1696	0.00	0.17
log $D_{7.4}$	1445	−0.01	0.21
rat hep CL_{int}	945	-0.07^a	0.26
rat PPB	668	−0.01	0.17
solubility (dried DMSO)	363	-0.08^a	0.36
solubility (solid)	466	-0.03^a	0.32

[a] Significantly different at the 95% level from zero.

The regression line in plot i of Figure 3 highlights that across the dynamic range of the human PPB response variable the typical stdev is approximately constant, or homoscedastic, as the slope for this line is not significantly different from zero at the 95% probability level. Therefore, the typical stdev for the response variable is given by the regression line intercept, which is approximately 0.17 on the $\log_{10}$(% bound/% free) scale, implying a gross experimental error criterion of a stdev greater than or equal to 0.34. In practice this criterion is very lenient; a molecule that had a measured human PPB of 98.00%, 99.00%, and 99.50% bound would have an acceptable stdev (i.e., ~0.30), even though there is a fourfold difference in the % free levels. Similarly, the gross experimental error criteria for log $D_{7.4}$ (plot ii of Figure 3) and rat PPB are greater than or equal to 0.42 and 0.34, respectively.

The regression line in plot iii of Figure 3 highlights that across the dynamic range of the solubility (dried DMSO) response variable, the typical stdev is not constant (i.e., heteroscedastic), as the slope for this line is significantly less than zero at the 95% probability level. This implies that the experimental error is larger for low-solubility molecules than for high-solubility molecules. For a molecule with three or more repeat measurements and a mean $\log_{10}$(solubility (dried DMSO)) equal to −1.00 (i.e., 0.10 μM) the typical stdev is 0.44. Similarly, for a molecule with three or more repeat measurements and a mean $\log_{10}$(solubility (dried DMSO)) equal to 3.00 (i.e., 1000.00 μM) the typical stdev is 0.12. Thus, the gross experimental error criteria will vary with the dynamic range, ranging from a stdev greater than or equal to 0.88 at the lower end to a stdev greater than or equal to 0.24 at the higher end. The plot for the solubility (solid) response variable is comparable as well as the three intrinsic clearance response variables. Such heteroscedasticity arises in early drug discovery assays because although they are assumed to be generic (for all possible molecules), in practice they are being pushed beyond their reliable capabilities. From plot iii of Figure 3, it is possible to argue that the reliable dynamic range for the solubility (dried DMSO) experimental data is ~1.00 μM to ~300.00 μM, within which the typical stdev is constant at ~0.30, rather than 0.10 to 1500.00 μM. From plot iv of Figure 3, it is possible to argue that the reliable lower limit for the human mic CL_{int} experimental data is ~25.00 μL/min/mg of protein, above which the typical stdev is constant at ~0.11, rather than 3.00 μL/min/mg of protein. The drive to be able to determine the solubility (dried DMSO) or human mic CL_{int} below 1.00 μM and 25.00 μL/min/mg of protein, respectively, comes at the expense of having, on balance, more variable data on which to build QSA(P)R models.

Systematic Experimental Errors. To create data sets of sufficient size and diversity to facilitate this study, measurements were pooled from different assays that were considered similar enough for each of the eight DMPK response variables. Experimental procedures deemed to have the potential of causing large systematic experimental errors have been considered. Where possible, these effects have been minimized by only considering assays that were similar on key procedural steps. However, a balance was required that accounted for the impact of removing such data against the need to maximize the data set sizes.

AstraZeneca has collected human PPB and rat PPB experimental data using either equilibrium dialysis or ultrafiltration in either neat or diluted plasma.[10,18] Only the data generated via equilibrium dialysis incubated at 37 °C was considered as this accounted for the majority; the ultrafiltration data tended to be generated at 25 °C and, as the position of the equilibrium is temperature dependent, systematic difference would be expected. Of the equilibrium dialysis data that could be considered, only that generated using neat plasma was included, eliminating the need to apply a single site binding assumption to the measurements generated in diluted plasma.[19] Log $D_{7.4}$ experimental data had been measured using either a shake flask or a retention time method.[11,20] Only shake flask data was considered as it provided a more generic method for lipophilicity determination, particularly for certain acidic molecules. It has been suggested that these acidic molecules make specific interactions with the chromatographic stationary phase that would lead to systematic differences.[21] Due to the potential effects of cosolvency on the octan-1-ol/water biphasic system, any log $D_{7.4}$ data generated from a DMSO liquid starting point was also excluded.[22] Solubility (dried DMSO) and solubility (solid) experimental data had been measured using either a centrifugation or filtration separation technique.[13] Only centrifugation data was considered due to the uncertainty associated with the filtration data regarding the potential nonspecific binding to the filter for hydrophobic molecules that would lead to systematic differences.

Of the remaining data sets used in this study, not all of the potential systematic experimental errors have been removed. These errors may include the use of: fresh and cryopreserved microsomes and hepatocytes, different CL_{int} incubation times, different rat plasma strains, different buffers, differences in room temperatures, dried DMSO starting material instead of solid, etc. However, it was assumed that these biases were small enough such that the magnitude of their effects could be tolerated within the study.

Random Errors. The typical random experimental error associated with a particular response variable measurement can be estimated by considering a large sample of stdevs associated with other molecules with repeat measurements. Figure 4 shows histograms of the percentage frequency distributions for the binned stdev seen for molecules with three or more repeat measurements for (i) human hep CL_{int} and (ii) solubility (solid) response variables. The distributions of these stdevs are not normal, but a property of the sampling distribution of the mean can be used to estimate the typical random experimental errors. The central limit theorem proposes that the distribution of all possible sample means (i.e., the sampling distribution of

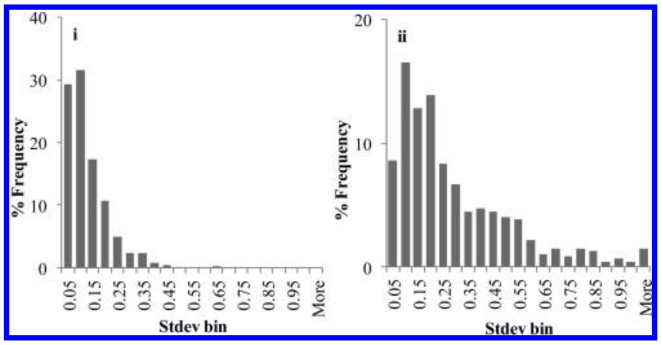


Figure 4. Percentage frequency distributions histograms for the binned stdev, for molecules with three or more repeat measurements, for (i) human hep CL_{int} and (ii) solubility (solid) response variables.

the mean), even if the original population is not normal, tends towards a normal distribution as the number of samples increases, and the mean thereof is the same as the population mean. By splitting the distributions into 50 random samples, the mean value of all the subsequent 50 sample means (drawn from these stdev distributions) can be used to estimate the true value (i.e., the typical stdev for the response variable); the stdev of this distribution of 50 sample means is an estimate of the standard error of the mean (i.e., the variability in the estimate of the typical stdev for the response variable). Table 3 summarizes the number of molecules with three or more repeat measurements, the range in the observed stdev, and the typical stdev for each response variable along with the 95% confidence limits for this estimate. The estimated typical stdevs are as low as 0.11 for human hep CL_{int} and as high as 0.28 for solubility (solid). As discussed earlier, the likelihood of building good QSA(P)R models is low if the experimental uncertainty in the response variable is typically greater than a stdev of 0.20. With the exception of the human hep CL_{int} and human mic CL_{int} data sets, the estimated typical experimental error for the six other DMPK response variables are such that good QSA(P)R models are unlikely.

HQ versus LQ Data. HQ data was defined as molecules with repeat measurements for a particular response variable with a stdev less than or equal to a certain criterion. Table 3 shows that for an individual molecule, the uncertainty in a single measurement may be exceptionally low (with a stdev of 0.01) or exceptionally high (with a stdev greater than 0.60), but for a large set of singly measured molecules it could be estimated by the typical stdevs for each response variable. However, for the purposes of building the QSA(P)R models discussed in the QSA(P)R Models—Part 1 section below, LQ data was defined as molecules with a single measurement as the uncertainty in the measurement could not be exactly specified.

A comparative study requires a comparison of the external test set predictions for a QSA(P)R model built on a training set of HQ data against those from a model built from a training set of LQ data, where the external test set has been drawn from the HQ data set. Furthermore, the predictor variable space for both the HQ and LQ data sets must be comparable. Analysis of the AstraZeneca data sets shows that 87% of the experimentally determined response variables for molecules were single measurements, and only 2% of the data sets had molecules with three or more repeat measurements; the remaining 11% had molecules with duplicate measurements. It was considered that there were not enough molecules with three or more repeat measurements to perform this study. Therefore, the definition for HQ data included molecules with two or more repeat measurements for a particular response variable with a stdev less than or equal to a certain criterion. The use of just two repeat measurements to estimate the population mean and stdev is itself quite poor, and is likely to impact on the ability to distinguish between HQ and LQ data; the use of a minimum of triplicate measurements would have been preferred but the AstraZeneca data sets did not permit this. The external test sets were randomly chosen from the HQ data sets using a minimum 80/20 split. Comparisons of the predictor variable space in the first two principal components indicate that the HQ data and LQ data sets occupied a similar space for each response variable; Figure 5 illustrates this for the log $D_{7.4}$ data sets.

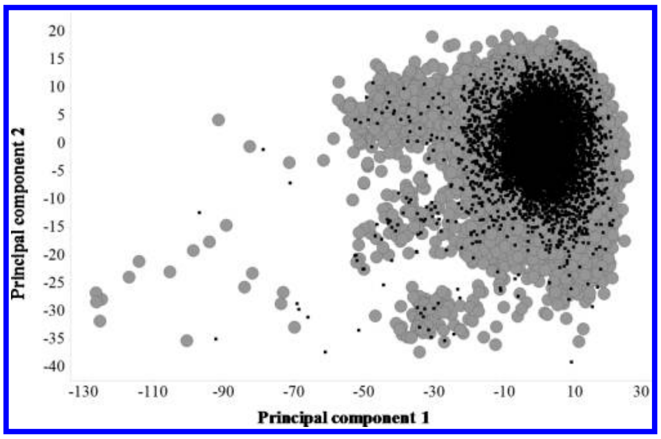


Figure 5. First two principal components for the HQ data (black squares) and LQ data (gray circles) for log $D_{7.4}$.

QSA(P)R Models—Part 1. Initially, models were built for each response variable using the HQ and LQ data sets, where

Table 3. Response Variable Summaries for the Distributions of Stdev, for Molecules with Three or More Repeat Measurements

response variable	number of molecules with three or more repeat measurements	range in observed stdev	typical stdev	lower 95% confidence limit for stdev	upper 95% confidence limit for stdev
human hep CL_{int}	540	0.01−0.61	0.11	0.05	0.17
human mic CL_{int}	830	0.01−0.67	0.12	0.08	0.16
human PPB	1696	0.01−1.56	0.16	0.11	0.21
log $D_{7.4}$	1445	0.01−2.12	0.19	0.11	0.27
rat hep CL_{int}	919	0.01−0.92	0.16	0.10	0.22
rat PPB	668	0.01−1.25	0.16	0.08	0.24
solubility (dried DMSO)	363	0.01−1.78	0.25	0.10	0.40
solubility (solid)	466	0.01−1.60	0.28	0.10	0.46

Table 4. Response Variable Summary for the HQ Data Models and the LQ Data Models

response variable	stdev criterion	RMSE training set HQ	RMSEP test set HQ	RMSE training set LQ	RMSEP test set LQ	Δ RMSEP	size of training set HQ	size of test set HQ	size of training set LQ	size of training set LQ/ size of training set HQ
human hep CL_{int}	0.10	0.15	0.34	0.11	0.45	−0.11	4196	988	3920	0.93
	0.20	0.14	0.34	0.11	0.43	−0.10	4610	1088	3920	0.85
	0.30	0.14	0.34	0.11	0.43	−0.09	4679	1106	3920	0.84
	0.40	0.14	0.35	0.11	0.44	−0.09	4702	1108	3920	0.83
human mic CL_{int}	0.10	0.16	0.37	0.15	0.44	−0.07	5774	1372	22132	3.83
	0.20	0.16	0.38	0.15	0.44	−0.06	6451	1537	22132	3.43
	0.30	0.16	0.37	0.15	0.44	−0.07	6646	1591	22132	3.33
	0.40	0.16	0.38	0.15	0.44	−0.06	6707	1605	22132	3.30
human PPB	0.10	0.19	0.45	0.17	0.43	0.02	3060	750	34972	11.43
	0.20	0.18	0.42	0.17	0.45	−0.02	4592	1084	34972	7.62
	0.30	0.18	0.43	0.17	0.46	−0.04	5255	1240	34972	6.65
	0.40	0.18	0.44	0.17	0.47	−0.03	5531	1303	34972	6.32
log $D_{7.4}$	0.10	0.24	0.62	0.23	0.55	0.07	4295	1011	34968	8.14
	0.20	0.24	0.55	0.23	0.54	0.02	5665	1339	34968	6.17
	0.30	0.24	0.58	0.23	0.54	0.03	6284	1491	34968	5.56
	0.40	0.24	0.58	0.23	0.56	0.02	6548	1569	34968	5.34
rat hep CL_{int}	0.10	0.14	0.33	0.15	0.41	−0.07	5911	1403	27100	4.58
	0.20	0.14	0.36	0.15	0.41	−0.05	6618	1585	27100	4.09
	0.30	0.15	0.35	0.15	0.40	−0.05	6883	1655	27100	3.94
	0.40	0.15	0.35	0.15	0.40	−0.05	6992	1676	27100	3.88
rat PPB	0.10	0.20	0.52	0.18	0.47	0.05	1118	269	13161	11.77
	0.20	0.20	0.43	0.18	0.42	0.02	1676	406	13161	7.85
	0.30	0.19	0.45	0.18	0.44	0.01	1873	476	13161	7.03
	0.40	0.19	0.45	0.18	0.45	0.01	1973	498	13161	6.67
solubility (dried DMSO)	0.10	0.28	0.70	0.26	0.63	0.07	1144	278	34920	30.52
	0.20	0.28	0.68	0.26	0.62	0.07	1532	369	34920	22.79
	0.30	0.29	0.67	0.26	0.61	0.05	1748	433	34920	19.98
	0.40	0.28	0.72	0.26	0.67	0.05	1860	473	34920	18.77
solubility (solid)	0.10	0.30	0.71	0.27	0.60	0.12	775	171	34058	43.95
	0.20	0.30	0.70	0.27	0.61	0.09	1108	268	34058	30.74
	0.30	0.30	0.72	0.27	0.59	0.13	1265	314	34058	26.92
	0.40	0.29	0.73	0.27	0.60	0.13	1377	339	34058	24.73

Table 5. Response Variable Summary for the Approximately Equal Sized HQ and LQ Data Models

response variable	stdev criterion	RMSE training set HQ	RMSEP test set HQ	RMSE (average) training set LQ	RMSEP (average) test set LQ	Δ RMSEP	% improvement in RMSEP	size of training pool LQ/size of training set HQ
human hep CL_{int}	0.10	0.14	0.34	0.11	0.44	−0.10	23.0	0.9
human mic CL_{int}	0.10	0.16	0.38	0.16	0.46	−0.08	17.3	3.8
human PPB	0.30	0.18	0.43	0.20	0.52	−0.09	16.4	6.7
log $D_{7.4}$	0.20	0.24	0.58	0.20	0.64	−0.06	9.4	6.2
rat hep CL_{int}	0.10	0.14	0.34	0.16	0.42	−0.08	19.3	4.6
rat PPB	0.30	0.19	0.45	0.22	0.57	−0.12	21.2	7.0
solubility (dried DMSO)	0.30	0.28	0.67	0.30	0.73	−0.05	7.2	20.0
solubility (solid)	0.20	0.30	0.73	0.32	0.75	−0.02	3.3	30.7

the stdev criterion for the former was investigated at 0.10, 0.20, 0.30, and 0.40. Any HQ data that exceeded the stdev criteria was excluded. Table 4 summarizes the QSA(P)R models' statistics. Notably, the external test set RMSEP is close to ~0.30 for the intrinsic clearance response variables which suggests that these are potentially good QSA(P)R models based on the above definition. With the exception of the human hep CL_{int} data, the sizes of the LQ data model training sets are greater than those of the HQ data model training sets, ranging from ~3 times for the human mic CL_{int} data to ~44 times for the solubility (solid) data. Overall, the HQ data models performed similarly to the LQ data models as measured by the external test set RMSEP. The best HQ data models occur with a stdev criterion of less than or equal to 0.30. It appears that when the ratio of the size of the LQ data model training set to the HQ data model training set is less than ~7, the external test set RMSEP for the HQ data model tends to be less than that for the LQ data model, the exceptions being the log $D_{7.4}$, and rat PPB models. For the solubility models (dried DMSO and solid), this ratio, of the size of the LQ data model training set to the size of the HQ data model training set, ranges from 18.77 to 43.95; in all cases the external test set RMSEP for the LQ data model is better than that of the HQ data model. There is an emerging pattern that as this ratio decreases, preferably as the proportion of the HQ data set increases, then the difference between the external test set RMSEP for the LQ data model and the HQ data model decreases, up to a point when the HQ data model becomes better. This pattern is most likely reflective of the molecular diversity within the HQ data model training set, and subsequently the predictive limits of the HQ data model.

To account for the effect of the LQ data model training sets tending to be much larger than the HQ data model training sets, models were built for each response variable using approximately equal-sized HQ and LQ data sets. The HQ data stdev criteria were fixed at the most appropriate level for each response variable (see Table 4); consequently, the training sets did not exceed 6000 molecules, and the test sets did not exceed 1500 molecules. In all cases the external test set was approximately a quarter of the size of either the HQ data model or the LQ data model training sets. With the exception of the human hep CL_{int} data, the LQ data model training sets were randomly chosen and, to account for the greater diversity, this process was repeated 50 times and a new model built each time. The data in Table 5 summarizes the QSA(P)R models' statistics. With respect to the LQ data models, the average training set RMSE and external test set RMSEP from the 50 repeat models are reported. As the human hep CL_{int} data set had too few LQ data points to properly permit this exercise, the average results are essentially from the same model as opposed to 50 different models. The HQ data models performed better than the LQ data models as measured by the external test set RMSEP. The percentage improvements in the external test set RMSEP for the HQ data models over the LQ data models range from 3.3% to 23.0%. These improvements can be rationalized by arguing that the LQ data models were built on experimental data that proportionally had greater uncertainty than the HQ data models. Unlike the HQ data model training set (and the HQ data external test set), the uncertainty in the LQ data model training set is undefined, although the information in Table 3 can help to estimate the likely uncertainty. For human hep CL_{int}, human mic CL_{int}, and rat hep CL_{int}, which were among the most improved response variables, the stdev criteria used to define the HQ data sets were less than or equal to 0.10. As the typical stdevs for these response variables were greater than this (i.e., 0.11, 0.12, and 0.16, respectively), it is likely that greater than 50% of the LQ data training sets contained molecules with single measurements with associated stdevs greater than 0.10. A similar argument can be made for the solubility (solid) response variable, although the improvement seen was only 3.3%. For the other four response variables, although the stdev criteria were all greater than the typical stdev, it can be reasoned from Table 4 that the majority of the HQ data being used had stdevs less than the typical stdevs. Therefore, the LQ data model training set can be considered to be associated with more experimental uncertainty than the HQ data model training sets for all eight DMPK response variables.

Figure 6 shows a plot, irrespective of response variable, of the ratio between the training set pool size for the LQ data models, (i.e., the size of the available LQ data set) and the size of the HQ data model training sets against the percentage improvement in the external test set RMSEP. It is possible to fit an exponential curve to this data. It follows that the largest improvements are seen for the response variables with the larger proportion of HQ data. It can be predicted from this exponential curve that a ~16% QSA(P)R prediction improvement for an external test set can be achieved by using a HQ data model training set one-sixth the size of a LQ data model training set. This relationship is probably indicative of how representative the HQ data model training set is of the external test set. The smaller the ratio, the more likely it will be to find representative molecules within the HQ data model training set.

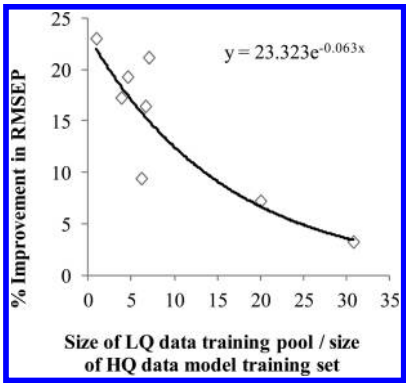


Figure 6. Ratio between the training set pool size for the LQ data and the size of the HQ data model training set, plotted against the percentage improvement in the external test set RMSEP for the HQ data model compared to the same sized LQ data models (see Table 5).

The LQ data has been defined as molecules with a single measurement and assumptions have been made about the associated experimental random error. Due to the nature of the data sets being studied, this permitted QSA(P)R models to be built on large data sets. However, it would be preferable to assign stdev criteria to both the HQ and LQ data definitions, although this would come at the expense of only being able to build QSA(P)R models on much smaller data sets.

QSA(P)R Models—Part 2. The definition of what constitutes LQ data was altered such that the HQ and LQ data were defined as molecules with repeat measurements for a particular response variable. For the HQ data, a stdev criterion of less than or equal to 0.10 was used. For the LQ data, a stdev criterion of greater than 0.10 was used. The HQ data model training set and external test set were randomly chosen from the HQ data set with a minimum split of 80/20. Models were built for each response variable using training sets of approximately equal size. For response variables where the HQ data model training set was smaller than the LQ data model training set, the LQ data model training set was randomly chosen to ensure that the two training set sizes approximately matched. This process of selecting training and test sets was repeated 50 times for each response variable and new models built each time. The data in Table 6 summarizes the QSA(P)R models' statistics, where the average values from the 50 repeat models are reported. The HQ data models performed better than the LQ data models as measured by the external test set RMSEP. The percentage improvements in the external test set RMSEP for the HQ data models over the LQ data models ranged from 1.3% to 26.0%. Student *t* tests (assuming unequal variances) on the separate distributions for the individual RMSEPs, calculated for the 50 HQ and LQ data models for each response variable, indicate that these percentage improvements were significant at the 99.9 level

Table 6. Response Variable Summary for the Approximately Equal-Sized HQ Data Models (Two or More Repeat Measurements with a Stdev Less than or Equal to 0.10), and LQ Data Models (Two or More Repeat Measurements with a Stdev Greater than 0.10)

response variable	RMSE training set HQ	RMSEP test set HQ	RMSE training set LQ	RMSEP test set LQ	size of training sets (average) HQ, LQ	size of test set HQ	Δ RMSEP	% improvement in RMSEP
human hep CL_{int}	0.16	0.39	0.12	0.49	638, 626	4546	−0.10	19.6
human mic CL_{int}	0.18	0.43	0.14	0.55	1249, 1233	5897	−0.12	21.6
human PPB	0.19	0.45	0.20	0.47	3049, 3054	761	−0.01	3.2
log $D_{7.4}$	0.25	0.62	0.27	0.63	3424, 3407	1882	−0.02	2.5
rat hep CL_{int}	0.15	0.37	0.15	0.50	1494, 1481	5820	−0.13	26.0
rat PPB	0.20	0.48	0.22	0.49	1110, 1107	277	−0.01	1.3
solubility (dried DMSO)	0.28	0.68	0.29	0.81	1139, 1137	283	−0.14	16.8
solubility (solid)	0.30	0.74	0.29	0.77	756, 754	190	−0.03	3.3

Table 7. Response Variable Summary for the Approximately Equal-Sized HQ Data Models (Two or More Repeat Measurements with a Stdev Less than or Equal to 0.10), and LQ Data Models (Two or More Repeat Measurements with a Stdev Greater than or Equal to 0.20)

response variable	RMSE training set HQ	RMSEP test set HQ	RMSE training set LQ	RMSEP test set LQ	size of training sets (average) HQ, LQ	size of test set HQ	Δ RMSEP	% improvement in RMSEP
human hep CL_{int}	0.17	0.42	0.10	0.55	147, 143	5037	−0.13	23.8
human mic CL_{int}	0.19	0.46	0.14	0.60	448, 443	6698	−0.14	23.2
human PPB	0.20	0.47	0.22	0.50	1790, 1780	2020	−0.03	5.5
log $D_{7.4}$	0.27	0.66	0.30	0.68	1817, 1800	3489	−0.02	3.3
rat hep CL_{int}	0.16	0.39	0.15	0.54	651, 642	6663	−0.15	27.5
rat PPB	0.21	0.51	0.23	0.54	572, 563	815	−0.02	4.5
solubility (dried DMSO)	0.29	0.70	0.29	0.91	774, 766	648	−0.21	23.1
solubility (solid)	0.30	0.74	0.29	0.77	750, 745	196	−0.03	4.4

except for the rat PPB models, where the improvement was only 1.3%. Table 7 shows that when the definition of LQ data used stdev criteria of greater than or equal to 0.20, this percentage improvement increased for all response variables, and ranged from 3.3% to 27.5%. Student *t* tests (assuming unequal variances) on the separate distributions for the individual RMSEPs, calculated for the 50 HQ and LQ data models for each response variable, indicate that these percentage improvements were significant at the 99.9 level. The range in percentage improvement is probably associated with the variation in the molecular diversity within each of the training sets. The RMSEP for the external test sets for all response variables in both scenarios, summarized in Tables 6 and 7, are all worse than those shown in Table 5. This can be explained by these models being built, on average, on ∼47% and ∼29%, respectively, of the size of the training sets in Table 5. The average relative size of the external test sets to the training sets for the scenario in Table 5 was ∼0.25, whereas for the scenarios summarized in Tables 6 and 7 it was ∼2 and ∼8, respectively. However, these percentage improvements can still be rationalized by arguing that the HQ data models were built on experimental data that had less uncertainty than that in the LQ data models.

It should be noted that these prediction improvements arise from applying stdev criteria filters to the experimental data and from QSA(P)R models generated using a random forest methodology. Of the eight DMPK response variables, none of the QSA(P)R models have an RMSEP for the external test set less than or equal to ∼0.30, so cannot be considered good according to the above definition. It is important to note that the HQ data sets studied contain molecules that have been chosen, for a variety of reasons, to have repeat measurements conducted on them. It is very unlikely that the principal reason for repeating these measurements was that these molecules were representative of the chemical space of interest for QSA(P)R modeling. Any overlap between the chemical space of the HQ data model training set and the external test set is fortuitous. It follows that prediction improvements may arise from having specifically designed, chemically representative HQ data model training sets and from investigating other modeling methodologies (e.g., support vector machines).

■ CONCLUSIONS

The routine collection of DMPK experimental data within early drug discovery can aid the evaluation of a molecule's potential as a drug and also guide the design of new molecules, which at times may be supplemented with QSA(P)R model predictions. Although the factors that determine whether a QSA(P)R model is good enough to aid subsequent design are subjective, it has been proposed that an RMSEP for an external test set of less than or equal to ∼0.30 is appropriate. No physical quantity can be measured with absolute certainty, so before any QSA(P)R model is generated, any gross and systematic experimental errors should be removed and the random experimental error quantified. When building QSA(P)R models, it is prudent to ensure that the random experimental error is defined by repeat measurements, and to only include those that pass a stdev criterion within the training set. Investigations of AstraZeneca internal data show that the predictive ability of QSA(P)R DMPK models improves as the random experimental error in the training set decreases. Specifically, training sets with measurements defined with stdevs of less than or equal to 0.10 are better than those with stdevs greater than or equal to 0.20, the caveat being that the molecular diversity within the training set needs to be similar to that of the molecules being predicted.

It follows that if a particular QSA(P)R model gave an RMSEP for an external test set of less than or equal to 0.30, it is possible to propose a change in the screening strategy for the particular response variable. Such a model would be defined as a good QSA(P)R model, allowing its predictions to be used for decision-making purposes, the model being able to predict a new molecule's true value within ±0.50 units on greater than or equal to 99% of the occasions. To put this into context, if the typical stdevs in Table 3 are used as the population stdevs of all possible measurements, the 99% confidence limits, (i.e., ± 2.58 × population stdev) on a single measurement would be greater than ±0.50 units for the solubility (dried DMSO and solid) response variables. Furthermore, if the population stdevs are more like the upper 95% limits, then only single measurements of human hep CL_{int} and human mic CL_{int} have 99% confidence limits for the true value that are less than ±0.50 units.

For the purpose of generating more influential QSA(P)R DMPK models, using experimental data from assays that can measure every new molecule is not recommended. Instead, assays that give rise to experimental data with a stdev of less than or equal to 0.20 on 99% of the occasions, from at least triplicate repeat measurements, are preferred. For early drug discovery DMPK assays there will clearly be a practical limit on the number of repeat measurements. However, it has been shown that QSA(P)R DMPK models need only be built on a representative one-sixth of all possible molecules, thereby reducing demand on experimental resources by potentially 50%.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44 1625 233103. E-mail: mark.wenlock@astrazeneca.com.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico DMPK/Tox: Why Models Fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83−92.

(2) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *THEOCHEM* **2003**, *62*, 39−51.

(3) Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25−26.

(4) Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297−4313.

(5) Miller, J. N.; Miller, J. C. Statistics of repeated measurements. In *Statistics and Chemometrics for Analytical Chemistry*, fourth ed.; Pearson Education Limited: Harlow, England, 2000; pp 20−41.

(6) Wood, D. J.; Buttar, D.; Cumming, J. G.; Davis, A. M.; Norinder, U.; Rodgers, S. L. Automated QSAR with a Hierarchy of Global and Local Models. *Mol. Inf.* **2011**, *30*, 960−972.

(7) Stålring, J. C.; Carlsson, L. A.; Almeida, P.; Boyer, S. AZOrange − High Performance Open Source Machine Learning for QSAR Modeling in a Graphical Programming Environment. *J. Cheminf.* **2011**, 3−28.

(8) Obrezanova, O.; Csányi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.

(9) Temesi, D. G.; Martin, S.; Smith, R.; Jones, C.; Middleton, B. High-Throughput Metabolic Stability Studies in Drug Discovery by Orthogonal Acceleration Time-of-Flight (OATOF) with Analogue-to-Digital Signal Capture (ADC). *Rapid Commun. Mass Spectrom.* **2010**, *24*, 1730−1736.

(10) Fessey, R. E.; Austin, R. P.; Barton, P.; Davis, A. M.; Wenlock, M. C. The Role of Plasma Protein Binding in Drug Discovery. In *Pharmacokinetic Profiling in Drug Research: Biological, Physicochemical, and Computational Strategies*; Testa, B., Kamer, S. D., Wunderli-Allenspach, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2006; pp 119−141.

(11) Wenlock, M. C.; Potter, T.; Barton, P.; Austin, R. P. A Method for Measuring the Lipophilicity of Compounds in Mixtures of 10. *J. Biomol. Screen.* **2011**, *16*, 348−355.

(12) Alelyunas, Y. W.; Liu, R.; Pelosi-Kilby, L.; Shen, C. Application of a Dried-DMSO Rapid Throughput 24-h Equilibrium Solubility in Advancing Discovery Candidates. *Eur. J. Pharm. Sci.* **2009**, *7*, 172−182.

(13) Wenlock, M. C.; Austin, R. P.; Potter, T.; Barton, P. A Highly Automated Assay for Determining the Aqueous Equilibrium Solubility of Drug Discovery Compounds. *J. Assoc. Lab. Autom.* **2011**, *16*, 276284.

(14) *Pipeline Pilot*, version 8.5; Accelrys: San Diego, CA, 2011.

(15) Strobl, C.; Malley, J.; Tutz, G. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol. Methods* **2009**, *14*, 32−348.

(16) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.

(17) *Dragon*, version 6.0.26; Talete srl: Milano, Italy, 2013.

(18) Wan, H.; Rehngren, M. High-Throughput Screening of Protein Binding by Equilibrium Dialysis Combined with Liquid Chromatography and Mass Spectrometry. *J. Chromatogr. A* **2006**, *1102*, 125−134.

(19) Berezhkovskiy, L. M. Consideration of the Linear Concentration Increase of the Unbound Drug Fraction in Plasma. *J. Pharm. Sci.* **2009**, *98*, 383−393.

(20) Valkó, K. Application of High-Performance Liquid Chromatography Based Measurements of Lipophilicity to Model Biological Distributions. *J. Chromatogr. A* **2004**, *1037*, 299−310.

(21) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F. ElogD_{oct}: A Tool for Lipophilicity Determination in Drug Discovery. 2. Basic and Neutral Compounds. *J. Med. Chem.* **2001**, *44*, 2490−2497.

(22) Li, A.; Yalkowsky, S. H. Predicting Cosolvency. 2. Correlation with Solvent Physicochemical Properties. *Ind. Eng. Chem. Res.* **1998**, *37*, 4476−4480.