

Locally Weighted Learning Methods for Predicting Dose-Dependent Toxicity with Application to the Human Maximum Recommended Daily Dose

Ruifeng Liu,* Gregory Tawa, and Anders Wallqvist

Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland 21702, United States

ABSTRACT: Toxicological experiments in animals are carried out to determine the type and severity of any potential toxic effect associated with a new lead compound. The collected data are then used to extrapolate the effects on humans and determine initial dose regimens for clinical trials. The underlying assumption is that the severity of the toxic effects in animals is correlated with that in humans. However, there is a general lack of toxic correlations across species. Thus, it is more advantageous to predict the toxicological effects of a compound on humans directly from the human toxicological data of related compounds. However, many popular quantitative structure–activity relationship (QSAR) methods that build a single global model by fitting all training data appear inappropriate for predicting toxicological effects of structurally diverse compounds because the observed toxicological effects may originate from very different and mostly unknown molecular mechanisms. In this article, we demonstrate, via application to the human maximum recommended daily dose data that locally weighted learning methods, such as *k*-nearest neighbors, are well suited for predicting toxicological effects of structurally diverse compounds. We also show that a significant flaw of the *k*-nearest neighbor method is that it always uses a constant number of nearest neighbors in making prediction for a target compound, irrespective of whether the nearest neighbors are structurally similar enough to the target compound to ensure that they share the same mechanism of action. To remedy this flaw, we proposed and implemented a variable number nearest neighbor method. The advantages of the variable number nearest neighbor method over other QSAR methods include (1) allowing more reliable predictions to be achieved by applying a tighter molecular distance threshold and (2) automatic detection for when a prediction should not be made because the compound is outside the applicable domain.

$$\text{Prediction} = \frac{\sum_{i=1}^p y_i e^{-(d_i/h)^2}}{\sum_{i=1}^p e^{-(d_i/h)^2}}$$

$$d_i \leq r_T$$

1. INTRODUCTION

The maximum recommended daily dose (MRDD) of a drug is an estimate of the upper daily dose limit above which the drug's efficacy is not improved, or its adverse effects outweigh the benefits.¹ For most drugs, the MRDD is limited by toxicity, and therefore, it is an important component of a drug's labeling for its safe use. Toxicity issues are also one of the main reasons why many drug candidates fail in development.² To reduce toxicity-related attrition in the later and more costly drug development stage, many structural moieties (alerts) known to be associated with an increasing chance of toxicity have been identified and used as filters in the earlier hit to lead discovery stage.³ Indeed, a recent analysis by Stepan et al. showed that the majority of drugs withdrawn from the market or required to have black box warnings because of unexpected toxicity contained at least one of these structural alerts.⁴ However, they also showed that about half of the top 200 drugs by prescription and sales in 2009 also contained at least one of the structural alerts. This indicates that one should not base go/no go decisions solely on structural alerts, as doing so would have prevented nearly half of the top-selling drugs from reaching the market and providing important medical benefits to the patient population. Instead, Stepan et al. found that a major differentiating factor between the top-selling drugs and those withdrawn from the market or

required to have black box warnings appeared to be the daily dose. Most of the top 200 drugs were administered at low daily doses in the range of milligrams to tens of milligrams, whereas most of the drugs withdrawn or having black box warnings were administered at daily doses exceeding several hundred milligrams.⁴ In this regard, being able to predict the MRDD would help assess the potential toxicity liability of a drug candidate and avoid drug development failures.

As pointed out by Contrera et al., the MRDD is essentially equivalent to the no observed adverse effect level (NOAEL) in toxicology.¹ In pharmaceutical development, animal NOAEL is used to estimate the starting dose of human clinical trials.⁵ The human equivalent dose (HED) is first constructed by converting animal NOAEL (in mg/kg-body wt/day), derived from toxicology studies, using the following equation:

$$\text{HED} = \text{NOAEL} \times [\text{animal body weight}(\text{in kg})/\text{human body weight}(\text{in kg})]^{0.33} \quad (1)$$

The HED is, in turn, converted to a maximum recommended starting dose (MRSD) by applying a safety factor, as follows:

Received: June 18, 2012

Published: September 10, 2012

$$\text{MRSD} = \text{HED}/\text{safety factor} \quad (2)$$

A default safety factor is 10, but values higher than 10 are used in cases where the drug has a steep dose–response curve, nonlinear pharmacokinetics, variable bioavailability, irreversible toxicity, or limited animal data. Safety factors lower than 10 are used when the drug under investigation belongs to a well-known compound class with well-characterized toxicity profiles.⁵ A fundamental assumption for estimating the starting dose in human clinical trials from animal NOAEL is that a compound's toxic effect on the animal is correlated with the toxic effect in humans. While this correlation may exist between humans and closely related species such as primates, it is much harder to recognize or justify such correlations between humans and more distant animal species, such as rodents, that are typically used in preliminary toxicological studies.⁶ As an example, Matthews et al. compared the human MRDDs of 326 compounds with the corresponding rodent maximum tolerated doses (MTD). The squared correlation coefficient between the $\log(\text{MRDD})$ and $\log(\text{MTD})$ of the 326 compounds was only ~ 0.20 , indicating less than desirable correlation between the two quantities.⁷

Given that MRDD is essentially equivalent to NOAEL, if we can predict the MRDD of a new compound with reasonable accuracy from the MRDDs of related compounds, the starting dose for a phase I clinical trial could be estimated directly by MRDD instead of NOAEL. This would significantly reduce the number of animals used and the expenses associated with preliminary toxicology studies. As pointed out by Contrera et al., the uncertainty and error associated with the MRSD derived from the quantitative structure–activity relationship (QSAR) modeling of human data will be less than the current animal extrapolation methods.¹ However, there have been surprisingly few publications on QSAR studies aimed at quantitative prediction of human MRDD, even though QSAR techniques are routinely used in many other aspects of predictive toxicology. This is partly explained by the fact that the toxic adverse effects of structurally diverse compounds originate from many different molecular mechanisms, e.g., hepatic toxicity due to reactive metabolites, cardiovascular toxicity due to the inhibition of K^+ channels, and central nervous system side effects due to compound penetration across the blood–brain barrier. It is not reasonable to expect a single global model derived from fitting all of the training set data to adequately predict the adverse effects.

Our literature search found only two QSAR studies for quantitative prediction of the MRDD. Both of them were from the Informatics and Computational Safety Analysis group of the United States Food and Drug Administration (FDA). The group collected MRDD data for 1,309 compounds from public sources such as The Physician's Desk Reference and Martindale: The Pharmacopoeia as well as from proprietary data submitted to the FDA. In the first study, molecular structures of the compounds were fragmented into 2- to 10-atom fragments. Statistical analyses were performed to identify fragments that might contribute to the adverse effects, and a numerical value was assigned to each fragment representing its contribution to MRDD. Prediction models were constructed based on the fragment contributions. An internal cross-validation using 120 compounds showed that the models were able to estimate MRDD to within a factor of 10 of the experimental values.⁷

In the second study, the MRDD values (in mg/kg-body wt/day) were converted into a logarithm activity scale.¹ To predict the MRDD for a target compound, a similarity search was first performed in the MRDD data set. Fifteen compounds most similar to the target compound and meeting a preset 80% similarity threshold were selected. A QSAR model was built correlating the MRDD of the 15 compounds with their two-dimensional molecular descriptors. The model was then used to predict MRDD for the target compound. In the study, molecular similarity was measured by the cosine coefficient or Tanimoto coefficient. Consistent with the observation that the Tanimoto coefficient is a more stringent similarity measure than the cosine coefficient,⁸ the authors found that for most of the compounds, there were <15 near neighbors in the 1,309-compound data set that met the 80% Tanimoto similarity criterion. In such situations, the cosine coefficient was used as the similarity measure. This approach for MRDD prediction is computationally demanding, as it requires the construction of a QSAR model, on the fly, for each compound for which a prediction was to be made. This individualized local QSAR model approach is compatible with the notion that structurally similar compounds are likely to exert their adverse effects via the same mechanisms. Internal validation using 120 compounds from the 1,309-compound data set indicated that the approach had 71% coverage, i.e., 29% of the compounds had <15 near neighbors meeting the 80% similarity criterion for a prediction model to be developed.¹

In essence, the method that the FDA group used in their second study is a k -nearest neighbor (k -NN) regression approach. Instead of performing explicit generalization for the whole data set, k -NN constructs hypotheses directly from the closest training examples. This method is most commonly used to classify objects,⁹ but it can also be used for regression.¹⁰ In the simplest form, k -NN assigns the average values of k nearest neighbors as the predicted value of the object for which a prediction is to be made. Significant improvements can be achieved by weighting the contributions of the neighbors so that the close neighbors contribute more to the prediction. In this sense, k -NN is a locally weighted learning method.¹¹ Locally weighted learning is well suited for modeling biological response induced by small molecules, as structurally similar compounds tend to have similar biological responses.¹² In recent years, we have seen more applications of k -NN in the computational study of molecular biological activity.^{13–18} In this article, we report results of our recent study applying both constant and variable numbers of nearest neighbors in locally weighted learning to predict human MRDD. We show that, in agreement with Contrera et al., the uncertainty and errors associated with the MRSD derived from locally weighted modeling of human MRDD are less than the current animal extrapolation methods. The same approach can also be applied to predicting/evaluating dose-related toxicity in animals and potentially reduce animal usage and expenses in the toxicological characterization of drug candidates.

2. MATERIALS AND METHODS

2.1. Methods. **2.1.1. Distance-Weighted k -NN Method.** Suppose we want to predict the biological activity y of a compound, and this compound has k nearest neighbors in the training set with activity values of y_1, y_2, \dots, y_k . A reasonable estimate of y would be the following:

$$y = \frac{\sum_{i=1}^k y_i}{k} \quad (3)$$

This estimate minimizes the squared error C , which is defined as follows:

$$C = \sum_{i=1}^k (y - y_i)^2 \quad (4)$$

However, this estimate treats all training data points equally, which is not consistent with the observation that structurally similar compounds are more likely to have similar bioactivity. A remedy is to apply a distance-related weight to the error criterion, as follows:

$$C = \sum_{i=1}^k [(y - y_i)^2 K(d_i)] \quad (5)$$

where d_i is the distance from a training set compound i to the query compound for which a prediction is to be made, and K is a weighting function. The best estimate should minimize C , i.e., $(\partial C)/(\partial y) = 0$, which is commensurate with the following:

$$y = \frac{\sum_{i=1}^k y_i K(d_i)}{\sum_{i=1}^k K(d_i)} \quad (6)$$

Depending on the numerical representation of molecular structures, the distance between two molecules can be defined in many ways. For example, with a conventional molecular descriptor representation, the Euclidean distance, $d(p, q)$, is defined as follows:

$$d(p, q) = \sqrt{\sum_{i=1}^m (q_i - p_i)^2} \quad (7)$$

where q_i and p_i are the corresponding values of descriptor i of molecules p and q , respectively. The cosine distance is defined as follows:

$$d(p, q) = 1 - \frac{\sum_{i=1}^m p_i \times q_i}{\sqrt{\sum_{i=1}^m (p_i)^2} \times \sqrt{\sum_{i=1}^m (q_i)^2}} \quad (8)$$

and the Tanimoto distance is defined as follows:

$$d(p, q) = 1 - \frac{\sum_{i=1}^m p_i \times q_i}{\sum_{i=1}^m (p_i)^2 + \sum_{i=1}^m (q_i)^2 - \sum_{i=1}^m p_i \times q_i} \quad (9)$$

For small molecules, it has been shown that the best measure of molecular similarity is the Tanimoto distance using molecular fingerprints as descriptors.⁸ The fingerprint-based Tanimoto distance, d_{Tan} , is defined as follows:

$$d_{\text{Tan}} = 1 - \frac{n(P \cap Q)}{n(P) + n(Q) - n(P \cap Q)} \quad (10)$$

where $n(P \cap Q)$ is the number of common on-bits in molecules p and q , and $n(P)$ and $n(Q)$ are the total number of on-bits of molecules p and q , respectively.

There are many reasonable weighting functions, such as the inverse distance, $1/d$, or the inverse distance squared, $1/d^2$. Here, we use a Gaussian kernel as the weighting function in the following k -NN procedure:

$$K(d) = \exp[-(d/h)^2] \quad (11)$$

where h is a smoothing factor. The appropriate value of h for a particular data set is determined via a cross-validation procedure.

In the k -NN approach, the number of appropriate nearest neighbors, k , is typically determined from cross-validation studies. Once a value for k is selected, it will remain the same for future predictions. In this sense, the conventional k -NN method is a constant number of nearest neighbor method.

2.1.2. Distance-Weighted Variable Number Nearest Neighbor Method. As pointed out by Contrera et al., Tanimoto similarity is a stringent molecular similarity measure. Not many compounds in the 1,309-compound MRDD data set have 15 or more neighbors meeting the $\geq 80\%$ Tanimoto similarity threshold (equivalent to $d_{\text{Tan}} = 0.20$). In order to use a constant number of 15 nearest neighbors to build their prediction models, Contrera et al. switched the similarity criterion to the less stringent 80% cosine similarity. In our opinion, an alternative and perhaps more meaningful approach is to use a consistent molecular similarity criterion, which predisposes the included compounds to exert their biological activity via the same molecular mechanism. This will result in a variable number of near neighbors, but a weighted average of the activity values of all the near neighbors should be a good estimate of the activity of the query compound. We call this the variable number nearest neighbor (ν -NN) method. Specifically, in the ν -NN method:

$$y = \frac{\sum_{i=1}^{\nu} y_i K(d_i)}{\sum_{i=1}^{\nu} K(d_i)} \quad (12)$$

where ν represents all the compounds meeting a molecular fingerprint Tanimoto distance threshold (TDT), and $K(d_i)$ is given by eq 11. Both the TDT and the smoothing factor are determined via cross-validation using training set compounds.

In this study, we evaluated the performance of both k -NN and ν -NN in predicting human MRDD. We used the Pipeline Pilot implementation¹⁴ of k -NN without modification, and we implemented ν -NN as described above in Pipeline Pilot, version 8.0.1.¹⁹ Functional class extended connectivity fingerprints with a diameter of four chemical bonds (FCFP_4)²⁰ were used as molecular descriptors in this study. For comparison with the performance of other commonly used regression methods, we also used a kernel-based partial least-squares (PLS) regression method implemented in Pipeline Pilot and support vector machine (SVM) regression as implemented in R.²¹ Molecular descriptors used by PLS and SVM regressions are AlogP,²² molecular weight (MW), and electrotopological-state atom type counts (E-State_count).²³ We also used E-State keys as descriptors, but a comparison of the results indicated that the performances of E-State counts and E-State keys were very similar, with the former marginally better when used with PLS and SVM.

2.2. MRDD Data Sets. **2.2.1. MRDD Training Set.** We used the FDA publicly disclosed MRDD database²⁴ containing a total of 1,220 entries as the training set. As described by the FDA group,^{1,7} most of the compounds in the data set are small organic molecule drugs. Organometallics, high-MW polymers ($>5,000$ Da), nonorganic chemicals (e.g., fibers, salts, and gases), mixtures of organic chemicals, and very small molecules (<100 Da) were excluded from the data set. The MRDD values are mostly oral single daily dose or its equivalent for an average adult (60 kg body wt).

2.2.2. MRDD external validation set. The FDA group also collated an additional group of 160 external validation compounds¹ that were not part of the training set from public sources and FDA files; 14 of the 160 compounds were proprietary, and their names and molecular structures were not disclosed. We retrieved molecular structure information for the rest of the compounds by their names from Drugbank²⁵ and PubChem²⁶ and used them as an external validation set in this study.

2.2.3. Conversion of MRDD Dose Units for QSAR Modeling. While the most prevalent dose unit for drugs, mg/kg-body weight, is easily understood by physicians, pharmacists, and patients, it is an inconvenient unit for modeling the relationship between an applied dose and the pharmacological response. The most common approach to study the relationship is to express the pharmacological response as a function of the logarithm of the applied dose, usually in molar concentration or mol/kg-body wt. In the present study, we converted the mg/kg-body wt/day doses into $\log(\text{mol/kg-body wt/day})$ via the molecular weight of the drugs.

2.2.4. Standardizing Molecular Structures for QSAR Modeling. For improved patient compliance, longer shelf life, cheaper storage and transportation, and improved aqueous solubility, pharmacologically

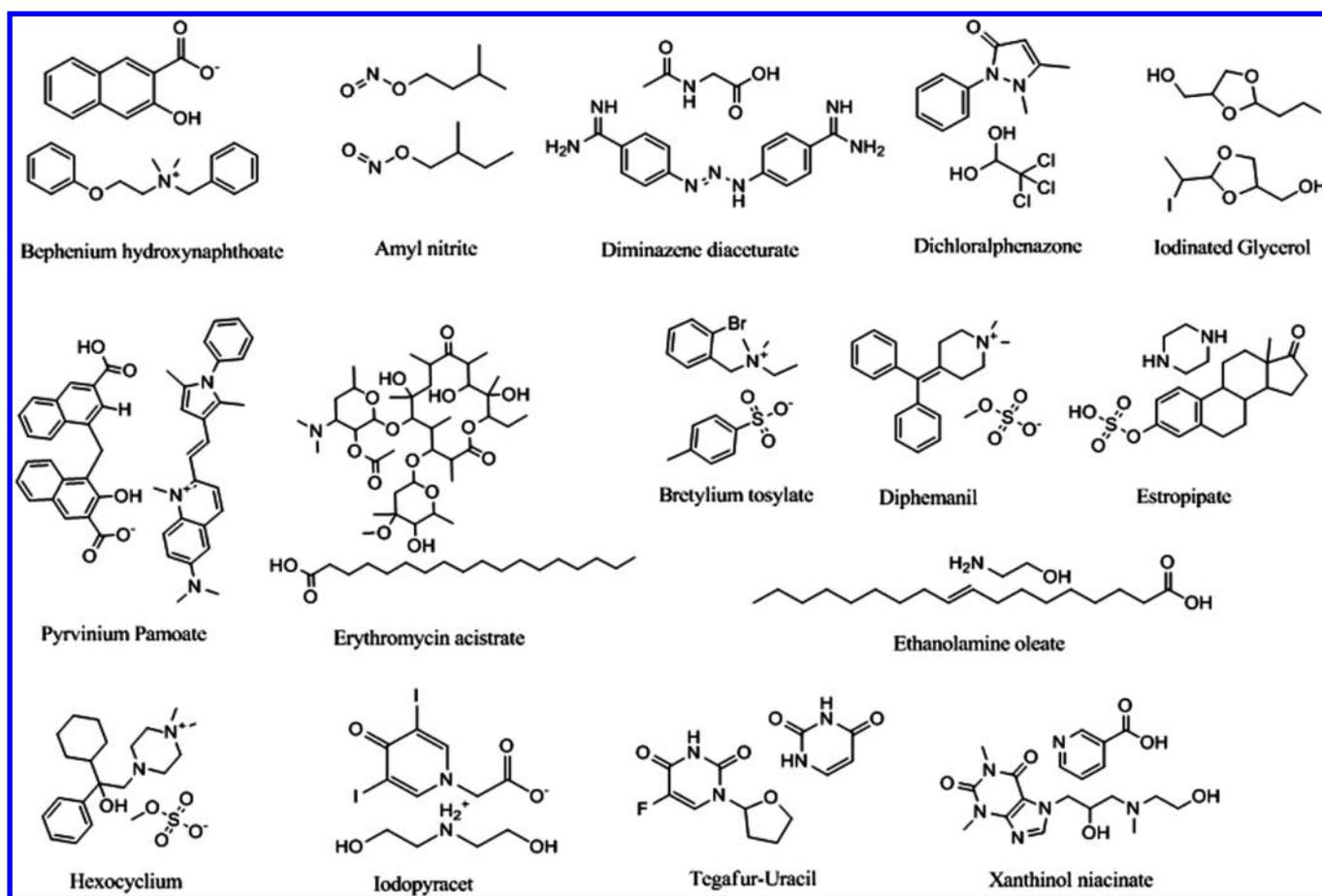


Figure 1. Molecular structures of two-component drugs in the FDA data set. They were excluded in the present study because of uncertainty in the dose-limiting components.

Table 1. Enantiomeric Pairs and Their MRDD in the FDA Data Set

| enantiomer pair | drug name | MRDD (mg/kg) | enantiomer pair | drug name | MRDD (mg/kg) | enantiomer pair | drug name | MRDD (mg/kg) |
|-----------------|----------------------------|--------------|-----------------|-----------------|--------------|-----------------|----------------------|--------------|
| 1 | dexbrompheniramine | 0.8 | 6 | lactose | 100 | 11 | pseudoephedrine | 4 |
| | brompheniramine | 0.4 | | maltose | 100 | | ephedrine | 1.67 |
| 2 | dextroamphetamine | 1 | 7 | levocarnitine | 49.5 | 12 | methadyl acetate | 2.33 |
| | amphetamine | 1 | | carnitine | 16.7 | | levomethadyl acetate | 4 |
| 3 | dexamethasone | 0.15 | 8 | galactose | 100 | 13 | propoxyphene | 6.5 |
| | betamethasone | 0.15 | | dextrose | 100 | | levopropoxyphene | 1.67 |
| 4 | dexamethasone acetate | 0.15 | 9 | dibromomannitol | 4.17 | 14 | thyroxine | 0.005 |
| | betamethasone acetate | 0.15 | | dibromodulcitol | 4.87 | | levothyroxine | 0.00833 |
| 5 | dexamethasone dipropionate | 0.15 | 10 | epirubicin | 2.43 | 15 | sorbitol | 833 |
| | betamethasone dipropionate | 0.15 | | doxorubicin | 2.43 | | D-mannitol | 999 |

active compounds are usually combined with inactive additives to become marketed drugs in different formulations. For oral drugs, the most common formulations are tablets or capsules of the salts of the effective components. In most of the cases, the additives are nontoxic and pharmacologically inactive. To develop a QSAR model for MRDD prediction, we removed all additives in this study. We understand that in individual cases, a specific formulation may help modulate the bioavailability of a drug and therefore may have an impact on the MRDD. However, there was insufficient information in the FDA data set for us to make a statistically reliable analysis of the effect at this point. We therefore chose not to address the effect of the adjuvant and assumed that it was within the uncertainty of the MRDD of the data set.

After the inactive additives of the drugs were removed, the MRDD in mol/kg-body wt/day was adjusted accordingly to reflect the change in the molecular formula. For example, if the formulation of a monoprotic acid drug is an Al^{3+} salt, the MRDD in mol/kg-body wt/day of the acid was adjusted to three times the MRDD of the aluminum salt in mol/kg-body wt/day. We then standardized the structures of the resulting effective components of the drugs by protonating all the acidic groups and deprotonating all basic groups.

There are, however, some two-component drugs where the second components are not simple acids but contain more complex organic structures. It is likely that both the beneficial pharmacological effects and dose-limiting adverse effects of these drugs are due to one or both components of these drugs. Figure 1 shows the molecular structures of

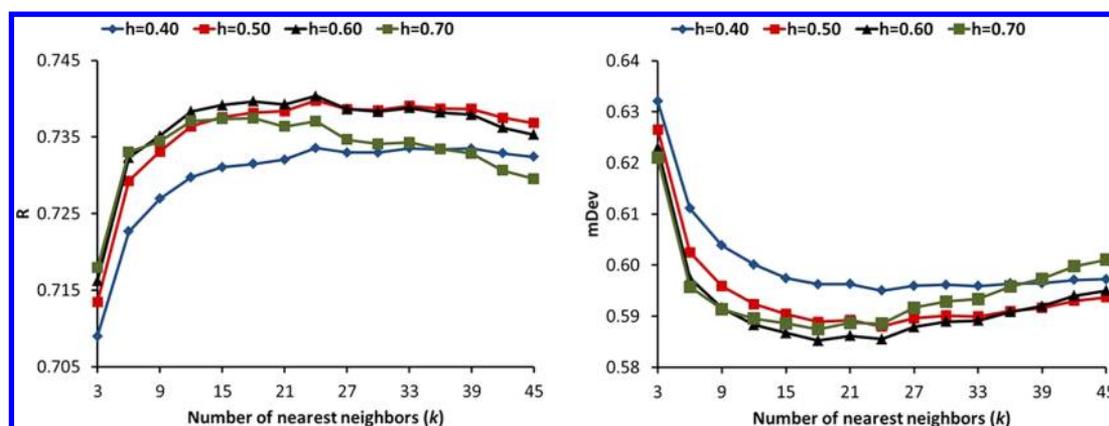


Figure 2. Performance of the constant number nearest neighbor (k -NN) method with different smoothing factors (h) and number of nearest neighbors, obtained from the 40-fold cross-validation of the training set compounds. R is the correlation coefficient, and $mDev$ is the mean deviation between k -NN predicted and database $\log(MRDD)$ values, where $MRDD$ is the human maximum recommended daily dose.

these drugs in the FDA data set. For the first six drugs, the FDA group listed both components as separate entries with identical $MRDD$ s. Since there was no information on which compounds are the dose-limiting components in these drugs, we decided to exclude them in this study. This reduced the total number of compounds in the training set to 1,199. In addition, there are 15 enantiomeric pairs of chiral compounds listed as separate entries in the FDA data set. Table 1 shows the compound names of these enantiomeric pairs and their $MRDD$ s. For each pair, the only difference in molecular structures of the two members is the chirality at their stereo centers. It is well-known that most drugs are chiral compounds, and in some cases, different enantiomers of a compound may have significantly different pharmacological effects.²⁷ However, most drugs on the market are racemic mixtures, presumably due to expenses in chiral synthesis/separation, and, in most cases, there are small differences in the efficacies. More importantly, the two-dimensional molecular structure-based approach we used is unable to account for the difference in molecular chirality. To properly investigate the impact of stereochemistry on $MRDD$, it is necessary to model three-dimensional molecular interactions between the drug and its biological targets that contribute to the desired pharmacological effect as well as to those targets that are involved in the dose-limiting adverse effects. Currently, there is insufficient information on the full complement of biological targets with which the drug molecules may interact. We therefore decided not to further consider the impact of molecular chirality on $MRDD$ in the present study. Table 1 shows that for each enantiomeric pair, the $MRDD$ values of the two members are very close, with the largest difference within a factor of 4 to each other. We retained one structure from each pair with the lower $MRDD$ value, whereas the structure with higher $MRDD$ was excluded from the training set. This reduced the number of compounds in the training set to 1,184.

The same structure standardization procedure was also applied to the external validation set compounds. One of the drugs, capreomycin, was excluded as it is a two-component drug with no clear indication as to which component or whether both components contribute to the dose-limiting adverse effects. This reduced the total number of our external validation set compounds to 145.

2.2.5. Details of QSAR Model Performance Evaluation. We evaluated the performance of the QSAR approaches for modeling $MRDD$ in two steps. The first was a 40-fold internal cross-validation using only the training set compounds. To do this, we randomly divided the training set compounds into 40 groups of ~ 30 compounds each. $MRDD$ values of all compounds in one group were predicted using the rest of the compounds in the data set as the training set. The process was repeated for every group so that the $MRDD$ value of every compound was predicted once. We then compared the predicted and database $MRDD$ values to determine the optimum number of nearest neighbors in the k -NN approach and the optimum TDT in the ν -NN approach. Appropriate smoothing factors were also evaluated by cross-

validation. The same 40-fold cross-validation procedure was also applied to the PLS and SVM methods.

In the second step of evaluating the performance of the QSAR approaches, all compounds in the 1,184-member training set were used for predicting the $MRDD$ of the external validation set, with the optimum number of nearest neighbors in the k -NN approach and the optimum TDT of the ν -NN approach fixed to the values determined from the 40-fold cross-validation of the training set.

Finally, we merged the external validation set into the training set so that locally weighted approaches have larger domain applicability, as the combined $MRDD$ data set covers a larger portion of chemistry space.

3. RESULTS AND DISCUSSION

3.1. Performance of the k -NN Method. As described in Materials and Methods, there are two tunable parameters that may affect the performance of the k -NN method, the smoothing factor h and the constant number of nearest neighbors k . They are usually determined by cross-validation calculations. In our experience, within a reasonable range h has a small impact on the performance of the k -NN method using the molecular fingerprint-based Tanimoto distance. However, the impact is much higher when extreme h values were used. For instance, when h approaches zero, distance penalties for the near neighbors are amplified. However, when a large h is used, the distance penalty is reduced. This is easy to understand as the weight function is a function of distance over h . In the Pipeline Pilot implementation of k -NN, 0.50 is the default h value. To find the optimal combination of h and k for the $MRDD$ data set, we performed 40-fold cross-validation calculations with h ranging from 0.40 to 0.70 and k ranging from 3 to 45. Figure 2 shows the results of the cross-validation calculations. In Figure 2, R is the Pearson correlation coefficient,²⁸ and $mDev$ is the mean deviation between the k -NN predicted and database $\log(MRDD)$ values. It shows that the optimal combination of h and k for this data set was $h = 0.60$ and $k = 24$. This combination gave the best k -NN performance, with the highest R of 0.74 and lowest $mDev$ of 0.59 in log units. The results also show that the variation in R and $mDev$ due to the difference in h was within 0.01 for R and 0.02 log units for $mDev$. However, k had a much higher impact on model performance. Starting with $k = 3$, model performance improved with increasing k , reaching the best performance at $k = 24$. From then on, the model performance deteriorated slowly with increasing k .

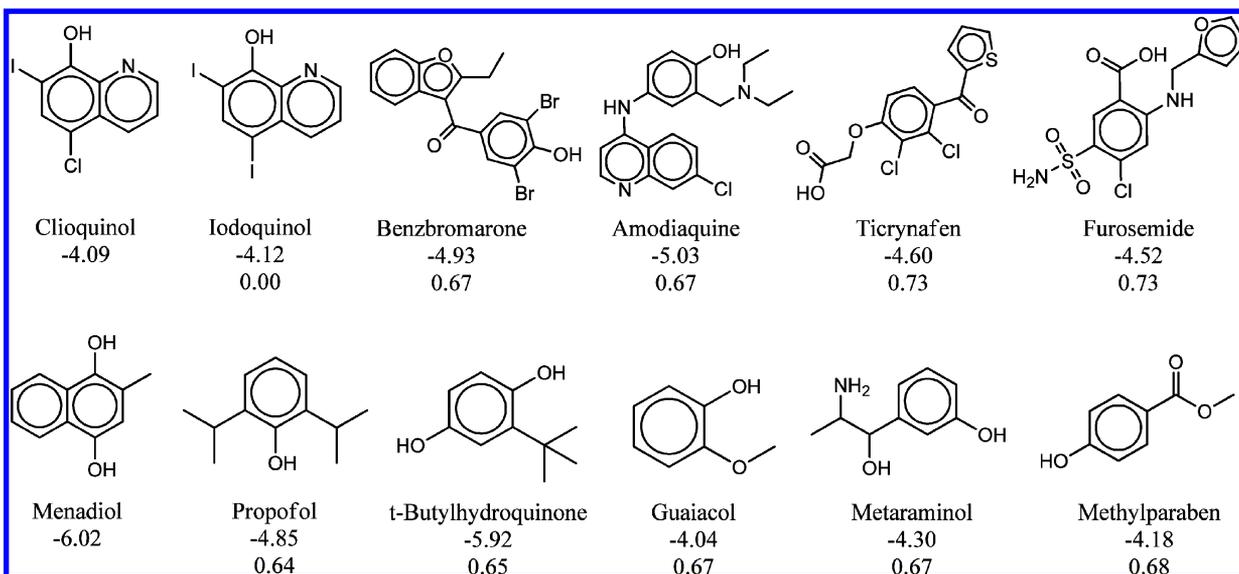


Figure 3. Example molecular structures of query compounds (clioquinol and menadiol) and their structurally closest neighbors in the training set. Top, clioquinol and its nearest neighbors; bottom, menadiol and its nearest neighbors. The *top* row of numerical values (negative numbers) are log(MRDD) values of each compound, and the *bottom* row of numerical values are Tanimoto distances of the compounds to their respective query compounds calculated with the FCFP₄ fingerprint.

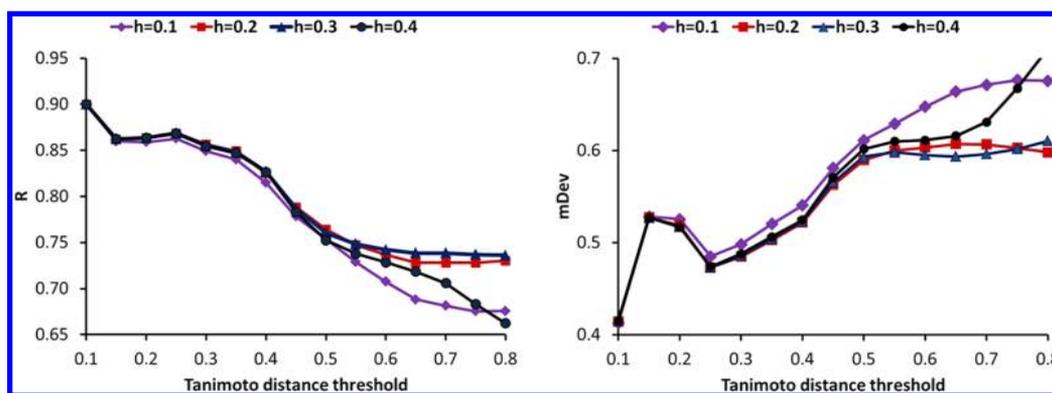


Figure 4. Performance of the variable number nearest neighbor (ν -NN) method with different h values and Tanimoto distance thresholds, obtained from 40-fold cross-validation of the training set compounds. R is the correlation coefficient, and $mDev$ is the mean deviation between ν -NN predicted and database log(MRDD) values, where MRDD is the human maximum recommended daily dose.

The inferior performance of the k -NN method with extremely small k may be due to large variations and uncertainties in the MRDDs for different therapeutic indications. Drugs used to treat acute infectious diseases may have high MRDDs, even at levels with significant adverse effects, because patients will be using them for very short time periods. Cancer drugs may also have MRDDs at levels with serious adverse effects. However, drugs for chronic use are understandably associated with low MRDDs. When k is too small, the predicted MRDDs are dominated by contributions from a very small number of training set compounds. With increasing k , increasing number of compounds contribute to the predictions via distance-weighted averaging, thus smoothing out large variations and uncertainties in the MRDDs of the training set compounds. This partly explains the slow deterioration of performance at increasing k after the optimal value of 24. With increasing k , increasingly dissimilar compounds are brought in the calculation as nearest neighbors. They should lead to inferior predictions because structurally dissimilar compounds are more likely to have different bioactivities. However, performance deterioration is limited

by distance-weighted averaging, resulting in the observed slow decrease in model performance as a function of increasingly large k .

In the k -NN method, it is assumed that a sample always has k nearest neighbors that are similar enough that they have the same mechanism of action. This, however, is hardly true for most molecular systems. Figure 3 shows some examples from the MRDD data set. In the first row of this figure, clioquinol has a single near neighbor, iodoquinol, which has a Tanimoto distance of 0.00 between them. The molecular structures of the two compounds are almost identical. It is very likely that they exert their adverse effect through the same mechanism(s), and, therefore, they have almost identical MRDD values. The rest of the compounds in the same row are the other closest near neighbors to clioquinol in the MRDD data set. However, since their Tanimoto distances to clioquinol are relatively large (0.67 or higher on a scale between close similarity at 0 to no similarity at 1), there is hardly any structural similarity to clioquinol. As a result, these compounds should not be considered as chemically close neighbors to clioquinol, and it makes no sense to include them in the k -NN prediction. The

second row of Figure 3 shows menadiol and its closest near neighbors in the data set. Since none of the compounds in the data set is structurally similar to menadiol, a MRDD prediction should not be made by k -NN for menadiol. However, by design, a prediction is always given by k -NN using the k nearest neighbors no matter how structurally similar the query is to its nearest neighbors. A better approach is to set a trusted molecular structural similarity threshold so that only compounds meeting the structural similarity criterion are included in the distance-weighted averaging. This is the ν -NN method described in Materials and Methods (section 2.1.2).

3.2. Performance of the ν -NN Method. We used the same 40-fold cross-validation procedure to tune h and the TDT of the ν -NN method as described in Materials and Methods (section 2.1.2). Figure 4 shows the results of the cross-validation calculations. As expected, with increasing TDT values, R decreased and mDev increased. The mDev data points at TDT values of 0.15 and 0.20 were outliers of the general trend. This is because with these extremely tight Tanimoto criteria, only a small number of compounds have qualified near neighbors in the data set. As a result, the predicted MRDD values are statistically less reliable. The situation was similar to the k -NN method with extremely low k values, as discussed in section 3.1.

Similar to the situation found in k -NN, within a reasonable h value range the performance of the ν -NN method was not sensitive to h (Figure 4). Since mDev increased with increasing TDT, the results indicate that one should use low TDT for better performance. However, the lower the TDT, the lower the number of compounds having qualified near neighbors and therefore the smaller is the applicable domain of the method. This is shown in Figure 5, where the results were derived from

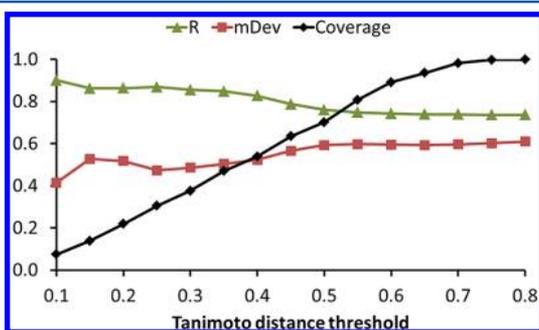


Figure 5. Performance and coverage of ν -NN obtained from 40-fold cross-validation of the training set with the h value fixed at 0.30.

the same 40-fold cross-validation procedure but with a fixed h value of 0.30. The coverage in Figure 5 is the percentage of compounds whose MRDDs can be predicted by the ν -NN method, i.e., the percentage of compounds with near neighbors meeting the TDT criteria. Obviously when applying the ν -NN method, one should choose a TDT by balancing prediction reliability and coverage of the method. With a large experimental data set, one can afford to use a tight TDT without sacrificing coverage because a query compound is more likely to have more qualified near neighbors in a large data set. For the FDA MRDD data set, a reasonable compromise was to set TDT to 0.45, which gave 64% coverage, an mDev of 0.56 log units, and an R value of 0.79 from the 40-fold cross-validation. By setting a molecular structure similarity threshold,

the ν -NN method can determine if a reliable prediction for a given compound can be made.

3.3. Comparison with PLS, SVM, and Nonfingerprint-Based k -NN Methods. PLS and SVM are perhaps the most popular regression methods in the QSAR community. The most common approach in QSAR studies using these methods is to develop a single global model to fit all the training set data. As discussed in the Introduction, this is not expected to work well for the MRDD data because multiple, and mostly unknown, mechanisms of action contribute to the dose-limiting adverse reactions. However, for comparison with the performance of the k -NN and ν -NN methods, we performed the same 40-fold cross-validation calculations using PLS and SVM. For PLS and SVM, FCFP_4 fingerprints are not suitable molecular descriptors because they only encode the presence or absence of a molecular fragment (e.g., a chlorine atom) and not how many times a molecular fragment is present in a molecule. The molecular fragment counts are crucial information for regression models. Because of this, we used AlogP, MW, and E-State counts as molecular descriptors.

The PLS calculation was done using Pipeline Pilot. A crucial parameter affecting PLS predictive power is the number of latent variables used. A higher number of latent variables used always improves the fit of the training set data but may lead to poorer prediction performance because of overfitting. Figure 6

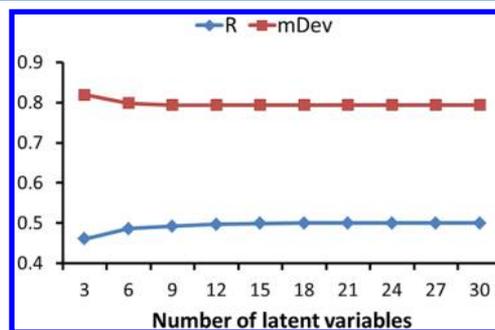


Figure 6. Performance of partial least-squares regression with different number of latent variables as measured by R and mDev values between the predicted and database $\log(\text{MRDD})$ values obtained from 40-fold cross-validation using the FDA MRDD training set.

shows R and mDev values obtained from 40-fold cross-validation using different numbers of latent variables in the PLS regression. As shown in Figure 6, the best performance that PLS can achieve for this data set is an R value of 0.50 and an mDev value of 0.79. These were compared with the corresponding best values derived from the fingerprint-based k -NN and ν -NN in Table 2, indicating inferior performance to the locally weighted methods.

We used the R implementation of SVM with a Gaussian radial kernel for performance comparison. The Gaussian radial kernel is a general-purpose kernel and usually performs better than other kernels, especially when there is no prior knowledge about the training data.²⁹ The option of epsilon regression was used to generate a regression model where epsilon was set to 0.1, indicating that the training set residuals were kept below 0.1 by increasing the number of support vectors as needed in the training process. There are two important parameters affecting the performance of nonlinear kernel-based SVM: the cost and gamma parameters. To find the optimal combination of the two parameters for the MRDD data set, we performed a 10-fold cross-validation with values of cost from 1 to 32 and

Table 2. Performance of Different Methods in 40-Fold Cross-Validation Using FDA Training Set Data

| method | descriptors | method parameters | R^a | mDev ^b |
|-----------|---------------------------|-------------------------|-------|-------------------|
| k -NN | ECFP_4 fingerprint | $h = 0.60, k = 24$ | 0.74 | 0.59 |
| ν -NN | ECFP_4 fingerprint | $h = 0.30,$ TDT=0.45 | 0.79 | 0.56 |
| PLS | AlogP, MW, E-State_Counts | 18 latent variables | 0.50 | 0.79 |
| SVM | AlogP, MW, E-State_Counts | cost = 4, $g = 0.05$ | 0.53 | 0.78 |
| k -NN | AlogP, MW, E-State_Counts | $h = 0.50, k = 90$ | 0.71 | 0.63 |

^aPearson's correlation coefficient between the predicted and database log(MRDD) values. ^bMean deviation between the predicted and database Log(MRDD) values.

values of gamma from 0.0125 to 0.2000. The optimal values were found to be 4 for cost and 0.0500 for gamma. These optimal parameters were then used in the 40-fold cross-validation as described above. The R and mDev values derived from the 40-fold cross-validation were 0.53 and 0.78, respectively. The data shown in Table 2 demonstrate that the performance of SVM was slightly better than PLS but worse than k -NN and ν -NN.

As both the k -NN and ν -NN calculations described above used molecular fingerprint as descriptors, we investigated how their performance compared with k -NN using conventional AlogP, MW, and E-State counts descriptors. Preliminary test calculations indicated that with the conventional molecular descriptors, the Euclidean distance worked slightly better than the Tanimoto distance for k -NN. We therefore performed the same 40-fold cross-validation using the Euclidean distance calculated from AlogP, MW, and E-State count descriptors in the k -NN approach. Figure 7 shows the results of the

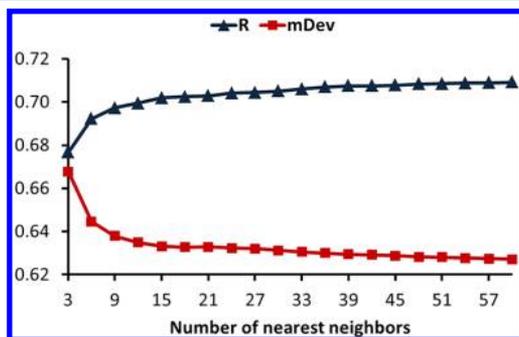


Figure 7. Performance of k -NN using AlogP, molecular weight, and electrotopological-state atom counts as molecular descriptors. The Euclidean distance given by eq 7 was used in the k -NN calculations.

calculations with increasing numbers of k and a fixed h of 0.50. It shows that performance of the k -NN improved slightly with increasing k , as R increases with increasing k , and a maximum value of 0.71 was achieved at a k of 90 with a corresponding mDev of 0.626 log units. They were close to the corresponding values of fingerprint-based k -NN and were significantly better than the corresponding values of PLS and SVM. This is consistent with our expectation that a single global model approach is not suitable for data sets encompassing biological responses from different molecular mechanisms.

3.4. Predictions for Compounds in the External Validation Set. To further examine the performance of the

fingerprint-based k -NN and ν -NN methods, we applied them to predict the MRDDs of the external validation set compounds. A total of 145 compounds survived our structure standardization steps. To make a prediction for a compound in the external validation set, a near neighbor search was performed in the 1,184-compound training set. In the k -NN prediction, the 24 closest training set neighbors were used irrespective of the actual Tanimoto distances between the external validation set compound and the 24 nearest neighbors in the training set. In the ν -NN approach, the TDT was set at 0.45 so that all training set compounds within 0.45 Tanimoto distance to the external validation set compound were used for prediction. If no compounds in the training set were within 0.45 Tanimoto distance to a compound in the external validation set, ν -NN did not make a prediction for the external validation compound. An h value of 0.60 was used for k -NN, and 0.30 was used for ν -NN as they were the optimal values from our 40-fold cross-validation experiments. Sixty of the 145 external validation set compounds did not have near neighbors in the training set within a Tanimoto distance of 0.45, and consequently, the ν -NN method did not make predictions for them. For the other 85 external validation set compounds, Figure 8 shows a

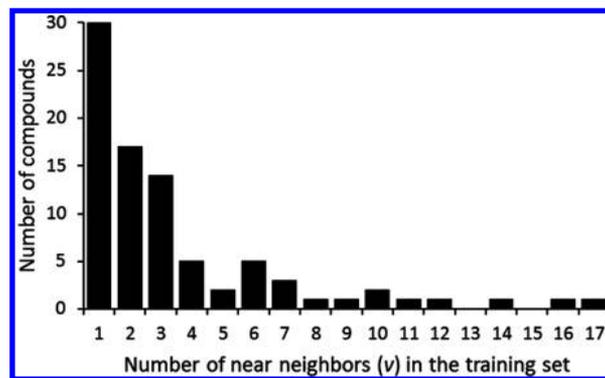


Figure 8. An overview of similarity between the training and external validation set compounds. The horizontal axis is the number of near neighbors (within a Tanimoto distance of 0.45) the external validation set compounds have in the training set. Bar height is the number of training set near neighbors the external validation compounds have. Not plotted is the bar height of validation set compounds having zero near neighbors, which is 60.

histogram of the number of training set near neighbors (within 0.45 Tanimoto distance). It shows that 30 of the external validation set compounds have only one training set near neighbor. Seventeen of them have two near neighbors in the training set. The number of validation set compounds having a higher number of training set near neighbors decreases significantly with increasing number of near neighbors.

Figures 9 and 10 show the results of MRDD predictions of the k -NN and ν -NN methods, respectively. For the 145 compounds, the R and mDev values of the k -NN predictions were 0.52 and 0.75, respectively. ν -NN gave predictions for 85 of the 145 compounds (59% coverage), with R and mDev values of 0.66 and 0.68, respectively. To illustrate the impact of the applicability domain, we also made ν -NN predictions for the 60 compounds without training set near neighbors within the 0.45 Tanimoto distance. This was done by using a large TDT of 1.0 so that all training set compounds were used in distance-weighted averaging. The results are presented in Figure 9 as red triangles. For the 60 compounds outside the

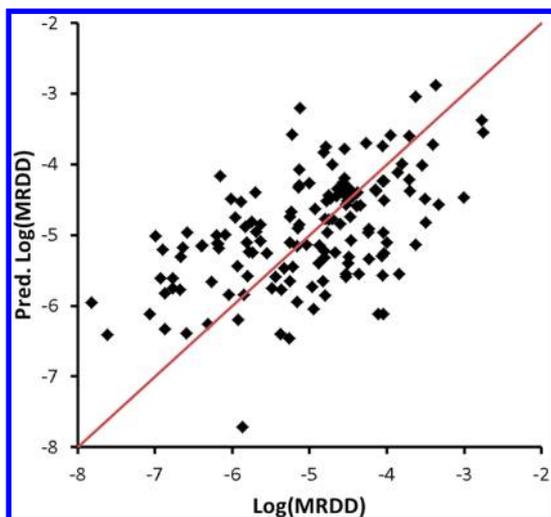


Figure 9. Correlation between the k -NN predicted and FDA external validation set $\log(\text{MRDD})$ values. The squared correlation coefficient was 0.27, and the mDev value was 0.75 log units. The calculations used the optimal values of k and h , 24 and 0.60, respectively, as determined from a 40-fold cross-validation of the training set compounds.

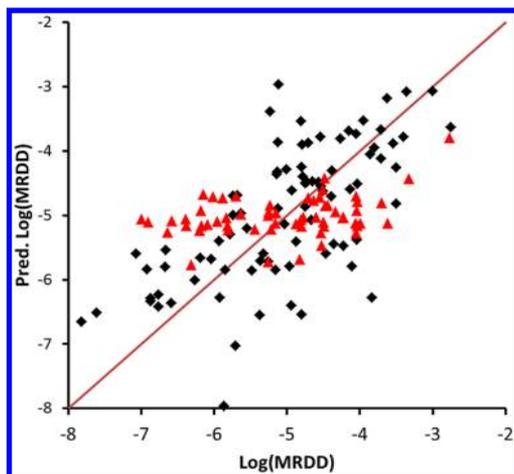


Figure 10. Correlation between the ν -NN predicted and FDA external validation set $\log(\text{MRDD})$ values. The black diamonds are results of ν -NN calculation using a TDT of 0.45. A total of 85 compounds in the external validation set have near neighbors in the training set within a Tanimoto distance of 0.45. For them, the squared correlation coefficient is 0.44, and the mean deviation is 0.68 log units. The red triangles represent external validation set compounds without near neighbors in the training set within 0.45 Tanimoto distance (outside of the applicability domain). Their predicted values were obtained using a TDT of 1.0 (all training set compounds were used in making the predictions). For the 60 compounds outside the applicability domain, the squared correlation coefficient is 0.14, and the mean deviation is 0.79 log units.

applicability domain of the ν -NN method, the R and mDev values are 0.37 and 0.79, respectively, indicating inferior correlation between the predicted and FDA MRDD values compared to that of compounds within the applicability domain of the method.

The FDA group published their QSAR predicted MRDD values (in mg/kg-body wt/day) for 99 of the 160 external validation set compounds (62% coverage). For the rest of the compounds, their QSAR model either gave bad statistics or had no coverage (compounds outside the applicable domain of the

model), or the predicted values were too high or too low and were considered off scale.¹ The logarithm of their predicted values versus the corresponding $\log(\text{MRDD})$ are shown in Figure 11 for comparison. The R and mDev values of their

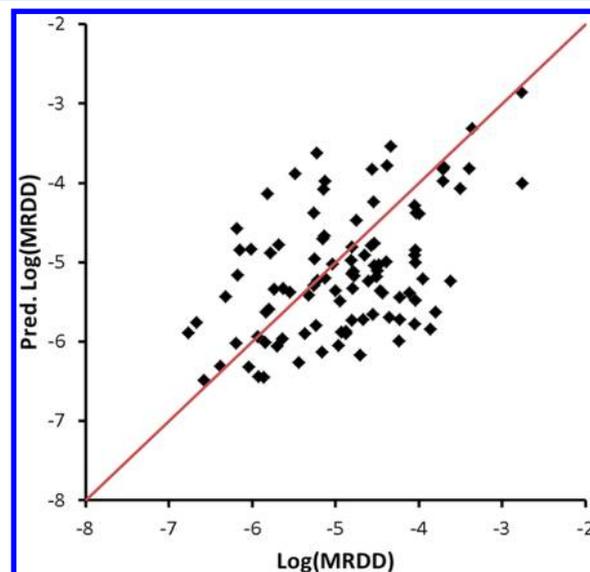


Figure 11. Correlation between the FDA QSAR predicted and external validation set $\log(\text{MRDD})$ values. The squared correlation coefficient was 0.23, and the mDev value was 0.70 log units.

predictions for the 99 compounds were 0.48 and 0.70, respectively. These values were similar to the corresponding values of the k -NN and ν -NN predictions. However, if one examines the squared correlation coefficients derived from the three methods (0.44, 0.27, and 0.23 of the ν -NN, k -NN, and FDA QSAR, respectively), it is clear that ν -NN performed the best, followed by k -NN and FDA QSAR. The mDev value between the predicted and experimental $\log(\text{MRDD})$ values does not appear to be a good indicator of method performance in this case because the FDA QSAR did not give predictions for most compounds with extremely low MRDD values. Figures 9 and 10 show that for compounds with extremely low MRDD values, the ν -NN and k -NN predictions also tended to have relatively large deviations.

4. CONCLUSIONS

In this study, we demonstrated that popular QSAR methods, which use a single global model to fit all training data, do not perform well in modeling the MRDD. The underlying reason for this poor performance was that the biological responses dictating the choice of MRDD arise from inherently different molecular mechanisms, and therefore, global chemical structure and structural similarity cannot be relied on to determine the biological response. However, locally weighted learning methods, such as k -NN, are well suited for such data sets. Locally weighted methods make predictions based on the closest near neighbors in the training set, and as long as the assumption of structurally similar compounds having similar bioactivities is valid, these methods naturally handle multiple toxicological mechanisms.

Because the k -NN method always uses a constant number of nearest neighbors in the training set to make a prediction for a query compound, regardless of whether the nearest neighbors are structurally similar enough to the query compound to

ensure similar bioactivity, it is expected that when a query compound lacks structural similarity to compounds in the training set, the k -NN prediction will be unreliable. This is the case when a query compound is outside the applicable domain of the training set. In this case, no prediction should be given by the method. The ν -NN method we introduced in the study is more advantageous than the k -NN method. In addition, ν -NN can give increasingly more reliable prediction by applying an increasingly tighter TDT. Because the bioactivities of structurally diverse compounds arise from multiple and usually unknown molecular mechanisms, locally weighted learning methods such as k -NN and ν -NN are better suited and should find more applications with time than methods relying on a single global model.

AUTHOR INFORMATION

Corresponding Author

*Phone: 301-619-1979. Fax: 301-619-1983. E-mail: RLiu@bhsai.org.

Funding

Funding for this research was provided by U.S. Department of Defense Threat Reduction Agency Grant TMTI0004_09_BH_T.

Notes

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Mr. Li Chen and Jason Smith for their technical assistance and the reviewers for their valuable suggestions.

ABBREVIATIONS

MRDD, maximum recommended daily dose; NOAEL, no observed adverse effect level; HED, human equivalent dose; MRSD, maximum recommended starting dose; MTD, maximum tolerated doses; QSAR, quantitative structure-activity relationship; k -NN, k -nearest neighbor method; ν -NN, variable number nearest neighbor method

REFERENCES

- (1) Contrera, J. F., Matthews, E. J., Kruhlak, N. L., and Benz, R. D. (2004) Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. *Regul. Toxicol. Pharmacol.* 40, 185–206.
- (2) Bombourg, N. (2011) Top R&D Drug Failures: Toxicity and Serious Adverse Events in Late Stage Drug Development are the Major Causes of Drug Failure. *GBI Market Research Report*, <http://www.marketresearch.com/GBI-Research-v3759/Drug-Failures-Toxicity-Serious-Adverse-6444690/> (accessed Aug 4, 2012).
- (3) Blagg, J. (2010) Structural Alerts for Toxicity, *Burger's Medicinal Chemistry. Drug Discovery and Development*, pp 1–34, John Wiley and Sons, New York.
- (4) Stephan, A. F., Walker, D. P., Bauman, J., Price, D. A., Baillie, T. A., Kalgutkar, A. S., and Aleo, M. D. (2011) Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* 24, 1345–1410.
- (5) Reagan-Shaw, S., Nihal, M., and Ahmad, N. (2008) Dose translation from animal to human studies revisited. *FASEB J.* 22, 659–661.
- (6) Barle, E. L., Looser, R., Cerne, M., and Bechter, R. (2012) The value of acute toxicity testing of pharmaceuticals for estimation of human response. *Regul. Toxicol. Pharmacol.* 62, 412–418.
- (7) Matthews, E. J., Kruhlak, N. L., Benz, R. D., and Contrera, J. F. (2004) Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Curr. Drug Discovery Technol.* 1, 61–76.
- (8) Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* 11, 1046–1053.
- (9) Weinberger, K. Q., and Saul, L. K. (2009) Distance metric learning for large margin nearest neighbor classification. *J. Machine Learn. Res.* 10, 207–224.
- (10) Rajagopalan, B., and Lall, U. (1999) A k -nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resour. Res.* 35, 3089–3101.
- (11) Atekeson, C. G., Moore, A. W., and Schaa, S. (1997) Locally Weighted Learning. *Artif. Intell. Rev.* 11, 11–73.
- (12) Wallqvist, A., Huang, R., Thanki, N., and Covell, D. G. (2006) Evaluating chemical structure similarity as an indicator of cellular growth inhibition. *J. Chem. Inf. Model.* 46, 430–437.
- (13) Li, H., Ung, C. Y., Yap, C. W., Xue, Y., Li, Z. R., Cao, Z. W., and Chen, Y. Z. (2005) Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem. Res. Toxicol.* 18, 1071–1080.
- (14) Jensen, B. F., Vind, C., Padkjaer, S. B., Brockhoff, P. B., and Refsgaard, H. H. (2007) In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k -nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* 50, 501–511.
- (15) Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., and Muller, K. R. (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* 49, 2077–2081.
- (16) Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Oberg, T., Dao, P., Cherkasov, A., and Tetko, I. V. (2008) Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* 48, 766–784.
- (17) Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A., and Tropsha, A. (2008) QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25, 1902–1914.
- (18) Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M., and Tropsha, A. (2009) Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* 22, 1913–1921.
- (19) Accelrys, Pipeline Pilot version 8.0.1, <http://accelrys.com/products/pipeline-pilot/> (accessed May 10, 2010).
- (20) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–54.
- (21) A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, <http://www.R-project.org> (accessed Aug 4, 2012).
- (22) Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of AlogP and ClogP methods. *J. Phys. Chem. A* 102, 3762–3772.
- (23) Kier, L. B., and Hall, L. H. (1990) An electrotopological-state index for atoms in molecules. *Pharm. Res.* 7, 801–807.
- (24) FDA Maximum Recommended Therapeutic Dose (MRTD) Database, <http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm092199.htm> (accessed May 10, 2012).
- (25) Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, 668–672.

(26) National Center for Biotechnology Information, PubChem Compound Database, <http://pubchem.ncbi.nlm.nih.gov/> (accessed May 10, 2012).

(27) Roden, D. M. (1994) Mirror, mirror on the wall ... stereochemistry in therapeutics. *Circulation* 89, 2451–2453.

(28) Adler, J., and Parmryd, I. (2010) Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry, Part A* 77, 733–742.

(29) Karatzoglou, A., Meyer, D., and Hornik, K. (2006) Support vector machines in R. *J. Stat. Software* 15, 1–28.