



Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities

Edwin J. Matthews^{a,*}, Carling J. Ursem^{a,b}, Naomi L. Kruhlak^a, R. Daniel Benz^a, David Aragonés Sabaté^c, Chihae Yang^d, Gilles Klopman^e, Joseph F. Contrera^f

^a US Food and Drug Administration, Center for Drug Evaluation and Research, Office of Pharmaceutical Science, Informatics and Computational Safety Analysis Staff (ICSAS), 10903 New Hampshire Ave., Silver Spring, MD 20993, USA

^b GlobalNet Services, 11820 Parklawn Drive, Rockville, MD 20852, USA

^c Prous Institute for Biomedical Research, S.A., Provenza 388, 08025 Barcelona, Spain

^d Leadscope, Inc., 1393 Dublin Road, Columbus, OH 43215, USA

^e Multicase, Inc., Ste 305, 23811 Chagrin Blvd., Beachwood, OH 44122, USA

^f Computational Toxicology Services LLC, P.O. Box 15665, Olney, MD 20830, USA

ARTICLE INFO

Article history:

Received 9 July 2008

Available online 30 January 2009

Keywords:

AERS
Computational toxicology
Drug adverse effects
Drug-induced liver injury
Drug-induced renal toxicity
In silico
Post-market reporting
Quantitative structure-activity relationships
QSAR software
SAR
SRS

ABSTRACT

This report describes the development of quantitative structure-activity relationship (QSAR) models for predicting rare drug-induced liver and urinary tract injury in humans based upon a database of post-marketing adverse effects (AEs) linked to ~1600 chemical structures. The models are based upon estimated population exposure using AE proportional reporting ratios. Models were constructed for 5 types of liver injury (liver enzyme disorders, cytotoxic injury, cholestasis and jaundice, bile duct disorders, gall bladder disorders) and 6 types of urinary tract injury (acute renal disorders, nephropathies, bladder disorders, kidney function tests, blood in urine, urolithiasis). Identical training data sets were configured for 4 QSAR programs (MC4PC, MDL-QSAR, BioEpisteme, and Predictive Data Miner). Model performance was optimized and was shown to be affected by the AE scoring method and the ratio of the number of active to inactive drugs. The best QSAR models exhibited an overall average 92.4% coverage, 86.5% specificity and 39.3% sensitivity. The 4 QSAR programs were demonstrated to be complementary and enhanced performance was obtained by combining predictions from 2 programs (average 78.4% specificity, 56.2% sensitivity). Consensus predictions resulted in better performance as judged by both internal and external validation experiments.

Published by Elsevier Inc.

1. Introduction

This is the second portion of a three-part investigation conducted by the US FDA's Center for Drug Evaluation and Research (CDER), Informatics and Computational Safety Analysis Staff (ICSAS). ICSAS is an applied regulatory research group that develops databases of toxicological and adverse human clinical information for use in data mining and quantitative structure-activity relationship (QSAR) modeling (Benz, 2007). In the first report we describe the creation of a human health effects database containing adverse event reporting data from two pharmaceutical post-market surveillance databases maintained by the FDA, the Spontaneous Reporting System (SRS) and the Adverse Event Reporting System (AERS), and from the published literature (Ursem et al., 2009). In

addition we described the method that was used to identify a subset of pharmaceuticals that had significant hepatobiliary and urinary tract adverse effects (AEs). In the current report we describe the creation of QSAR models to predict hepatobiliary and urinary tract AEs of drugs based upon the molecular structure of the pharmaceutical. We employed four state-of-the-art global QSAR software programs and report the experimental parameters and methods that were needed to optimize the predictive performance of these models. In the third report we describe specific properties of the drugs that had significant hepatobiliary and urinary tract AEs. These properties include both the clinical indication(s) for which the drug was approved, and the multiple pharmacological activities of the drug predicted by QSAR programs (Matthews et al., 2009). The overall goal of that investigation was to devise a generalized *in silico* methodology which could be utilized to predict AEs of pharmaceuticals and provide some insight into possible mechanisms of action (MOAs) responsible for the AEs.

* Corresponding author. Fax: +1 301 796 9998.

E-mail address: edwin.matthews@fda.hhs.gov (E.J. Matthews).

Nomenclature

Actives	a subset of drugs reported to cause AEs in patients (pharmaceutical) adverse effect	PRR	proportional reporting ratio (disproportional analysis of pharmaceutical AEs)
AE		QSAR	quantitative structure-activity relationship
AERS	FDA's Adverse Event Reporting System	RCCP	statistic: rate of change of the predictive performance of an AE QSAR model
BP	the breakpoint activity value that distinguishes active from inactive drugs	SRS	FDA's Spontaneous Reporting System
Inactives	a subset of drugs that had no significant AEs reported	SDF	structure-data file
LMO	leave-many-out (cross-validation experiment)	WOE	weight of evidence
MOA	mechanism of action		

FDA/CDER utilizes the post-market SRS and AERS systems to collect reports of pharmaceutical AEs from manufacturers, physicians and patients (<http://www.fda.gov/medwatch/SAFETY.htm>) and uses this information to monitor AEs that may not have been detected during pre-clinical animal testing and clinical trials. In general, clinical trials are more efficient at detecting beneficial pharmacological effects and dose-related organ toxicities than idiosyncratic adverse occurrences (Navarro and Senior, 2006). Post-market AE data are analyzed by pharmacovigilance groups to identify any significant drug-related AE signal that could be the subject of further study and analysis. Traditionally, a proportional reporting ratio (PRR) is used by pharmacovigilance groups to account for variations in AE reports due to different patient populations and exposures for each drug (Evans et al., 2001; Moore et al., 2005). Using the method described in these publications, a significant drug-related AE must have at least 3 observed AE reports, a χ^2 of 4, and a Yates correction of the χ^2 value.

Despite the potential utility of QSAR models to provide decision support information in drug discovery, lead selection, and regulatory issues, little work has been done to identify global structure-activity relationships (SAR) for drugs having toxicologically related AE findings. A previous attempt by this laboratory to compile and construct QSAR models using FDA's SRS post-market AE data showed some success (Matthews et al., 2004), but we suggested then that a larger database of observations would be required to produce models with greater predictive performance. This study represents an extension of our earlier effort.

An important objective of this investigation was to devise a generalized methodology which could unambiguously discriminate a subset of drugs that was associated with unexpected and serious AEs in patients (hereafter called actives) from drugs that had produced no significant AEs (hereafter called inactives). In addition, we required that this method would identify all actives that historically had been removed from the market because of AEs. Our operating hypothesis was that the active drugs that represented the highest risk of causing serious organ failure resulted in more than one type of AE, i.e., multi-AE endpoint activities, and the multiple AEs were presumably related to different mechanisms by which the drug injured an organ. It was further postulated that the active drugs with multi-AE activity for the same organ/system might share chemical molecular properties that could be recognized by QSAR software programs. Our hypothesis was inspired by the observations of Hyman Zimmerman (1978, 1999), who reported that drugs causing drug-induced liver injury by two different mechanisms in the same patient, i.e., an increase in a liver enzyme such as alanine or aspartate aminotransferase and also jaundice, are most likely to cause serious human liver toxicity in that patient. Our strategy for testing this hypothesis was to separate active and inactive drugs utilizing two independent experimental parameters: (1) identification of clusters of toxicologically related AE endpoints, and (2) identification of clusters of active drugs that shared molecular properties recognized by QSAR software pro-

grams. In the companion study we describe a statistical weight of evidence (WOE) method to organize a majority of hepatobiliary and urinary tract AEs into 11 AE clusters (Ursem et al., 2009). In this study we describe a statistical method to separate active and inactive drugs for each of the 11 AE clusters that is based upon the predictive performance of QSAR software programs. This approach is supported by our successful modeling of pre-clinical rodent carcinogenicity and reproductive toxicity by clustering the most toxic compounds that produced trans-gender and trans-species effects (Matthews et al., 2007a,b).

The second objective of this investigation was to construct a battery of QSAR models that could be used to predict drug-induced hepatobiliary and urinary tract injury in humans. We chose to investigate two different organ systems simultaneously in order to generalize our methodologies for identifying AEs for other organ systems. Hepatological and renal AEs are often rare and idiosyncratic, making them difficult to predict before a drug is marketed due to the limited statistical power of clinical trials. The liver was chosen as an important organ system for developing QSAR models because drugs are the single most common cause of acute liver failure (Lee, 2003, 2005) and hepatotoxicity is the most common reason for the FDA to take regulatory action against a drug (Temple, 2001). A study of all safety withdrawals of prescription drugs from worldwide markets from 1960 to 1999 (Fung et al., 2001) showed that internationally, the liver is the most common organ to be affected by toxicities leading to drug withdrawal, with 26.2% of all withdrawals being attributed to it. In recent years the FDA has partnered with members of academia and the pharmaceutical industry to form a workgroup focused solely on better understanding the causes and warning signs of drug-induced liver toxicity (<http://www.fda.gov/cder/livertox>). The urinary tract was also chosen for QSAR modeling because our laboratory's previous experience with this organ system (Matthews et al., 2004) suggested that a mechanism related to chemical structure appeared to be involved in renal toxicity, a theory also supported by the literature (Fielden et al., 2005).

The decision to employ four state-of-the-art QSAR software programs to model AEs was motivated by our experience in using consensus strategy to predict carcinogenicity in rodents (Contrera et al., 2007; Matthews et al., 2008), and by our desire to develop a generalized methodology for constructing AE QSAR models that can be applied to different QSAR programs with different prediction paradigms. We feel that the use of several QSAR programs and a consensus prediction strategy offers three main advantages over using a single QSAR program. (1) Whereas individual QSAR programs can be configured to provide high specificity predictions of chemical toxicity (>80%), the sensitivity is invariably low and dependent upon the QSAR program and toxicological endpoint. In our experience no one QSAR program provides both high specificity and high sensitivity predictions. However, higher sensitivity predictions with good specificity can be attained using a consensus prediction strategy. (2) The use of two or more QSAR programs

with different prediction paradigms under high specificity conditions could potentially provide complementary molecular property information on toxic drug molecules. For example, molecular fragment- and molecular descriptor-based paradigms can provide sub-molecular structural alerts and whole molecular descriptor properties highly correlated with chemical toxicity. (3) When multiple QSAR programs identify the same active molecules, there is a higher probability that there is a causal relationship between the active molecule and its structure. Actives identified by consensus prediction have a molecular property space that is uniquely different from inactive molecules and thereby offers a practicable way to investigate mechanisms of action of drugs and their AEs. This strategy is based upon the hypothesis that the best QSAR model predictive performance is obtained when the active and inactive drug molecules are correctly classified and optimally separated into two groups with the active molecules forming distinct clusters based upon shared chemical molecular properties, while poor predictive performance is hypothesized to be caused by misclassification of active and inactive molecules.

It is our hope that these QSAR models may be used to provide decision support information to pharmaceutical industry in drug discovery and lead chemical selection applications, as well as be valuable for FDA's regulatory and research activities.

2. Materials and methods

2.1. Pharmaceutical AEs and chemical structure databases

2.1.1. AE data sources and defined endpoints

The hepatobiliary and urinary tract organ system AE data compiled for this study and the companion report (Ursem et al., 2009) came from two pharmaceutical post-market surveillance databases maintained by the FDA, SRS (1969–1997) and AERS (1997–present), and from the published literature accessed through MicroMedex Integrated Index (<http://csi.micromedex.com>) and Elsevier's PharmaPendium (<http://www.pharmapendium.com>). The nomenclature for all of these AE data was standardized based upon the MedDRA terminology (<http://www.meddrasso.com>) for 229 hepatobiliary and 426 urinary tract specific “preferred term” AEs. Analyses of the data revealed that the majority of the specific AEs were associated with eleven composite MedDRA, “high level term” cluster AE endpoints. The hepatobiliary data consisted of 5 composite endpoints: liver enzyme disorders, cytotoxic injury, cholestasis and jaundice, bile duct disorders, and gall bladder disorders; the urinary tract consisted of 6 composite endpoints: acute renal disorders, nephropathies, bladder disorders, kidney function tests, blood in urine, and urolithiasis. The companion report describes in detail: (1) the methods used to compile these AE data, (2) the statistical analyses to identify clusters of specific AE endpoints that were highly correlated with one another, and (3) a WOE method to quantify and transform textual data into a numerical format suitable for QSAR modeling.

2.1.2. Estimation of patient exposure and significant AEs

The methods we used to estimate patient exposure and significant AEs are described in the companion paper (Ursem et al., 2009). Briefly, disproportionality analyses were performed that make use of a denominator reflecting patient exposure to a drug based upon the frequency of reporting for that drug in the entire database (Evans et al., 2001). The use of this denominator is based upon the assumption that in the absence of a toxicological mechanism causing an AE, AEs for drugs and endpoints occur independently. The actual number of reports (observed) divided by this expected value is defined as the Proportional Reporting Ratio (PRR), which is used to detect drug AE signals. To calculate the

expected value, the total numbers of reports in the entire database of 9685 human AE endpoints were summed for each generic drug and each AE endpoint. The fraction of total reports per drug multiplied by the total reports for each endpoint was then calculated to be the expected value for each drug at each specific endpoint. To calculate the observed value, all reports for each drug at each endpoint were pooled. When the observed value was greater than the expected value and resulted in a large PRR value that was statistically significant ($\chi^2 > 4$), a drug was considered active at that endpoint. In order to prioritize pharmaceutical AE findings for data mining and QSAR modeling, compounds were scored as either active (positive) or inactive (negative) by setting a breakpoint (BP) to distinguish active from inactive PRR values. The BP was set using several different PRR methods in order to determine the method that resulted in the best QSAR model predictive performance (Section 2.4).

2.1.3. Chemical structure database

Pharmaceutical chemical names and structures were standardized across the AERS and SRS databases as described in the companion paper (Ursem et al., 2009). For the purposes of QSAR modeling, AE data for the same drug listed under multiple trade names were pooled under a standardized drug generic name. Additionally, AE data were combined for generic drugs available as different salt and ester compounds. Chemical structures were saved as ISIS “.mol” files and as Simplified Molecular Input Line Entry System (SMILES) codes generated by MC4PC.

2.1.4. Applicability domain of the AE QSAR models

The ICSAS AE database only includes pharmaceuticals that have been used to treat human clinical indications. However, certain pharmaceuticals were excluded from the QSAR modeling because they cannot be analyzed by the QSAR software programs used in these studies: inorganic chemicals (salts and metals), high molecular weight substances (>5000 Da) (peptides, polysaccharides, proteins, fibers, etc.), organometallic chemicals, gases, and mixtures of chemicals. Combination drug product mixtures and high molecular weight substances account for approximately 50% of the drugs in our AE database. The applicability domain of the AE QSAR models was shown to be very high for the remaining pharmaceuticals considered in this study, and we anticipate that the models will show high coverage for other FDA regulated substances such as cosmetics and natural substances, but may demonstrate poorer coverage for industrial chemicals such as those regulated by the US Environmental Protection Agency.

2.2. QSAR software

This investigation utilized four software programs: (1) MC4PC (versions 1.5 and 1.7) provided by MultiCASE, Inc. (<http://www.multicase.com>); (2) BioEpisteme (version 2.0) provided by Prous Institute for Biomedical Research, S.A. (PIBR, <http://www.prouscience.com>); (3) MDL-QSAR (version 2.2) provided by MDL Information Systems, Inc. (<http://www.mdli.com>); and (4) Leadscope Predictive Data Miner (LPDM version 2.4) provided by Leadscope, Inc. (<http://www.leadscope.com>). All of the programs are machine learning tools that create global QSARs (Gombar et al., 2007; Matthews and Contrera, 2007). MC4PC, MDL-QSAR, and LPDM were obtained under Cooperative Research and Development Agreements (CRADAs) between FDA and the collaborating software developers, while BioEpisteme was obtained via a Material Transfer Agreement. All of the AE QSAR models described herein were developed by the FDA.

The same training data set was used to construct a QSAR model for each of the four programs. The AE activities of the drugs in the training data sets were characterized using a binary classification system in which drugs were classified either as being of low risk

of AEs (0; negative; inactive; or 10–29 activity units) or as high risk (1; positive; active; 30–80 activity units). The activity unit scoring was similar to that used in our rodent carcinogenicity (Matthews and Contrera, 1998); genetic toxicity (Matthews et al., 2006a,b); and reproductive and developmental toxicity (Matthews et al., 2007a,b) investigations. Even though the activity scores were converted to a binary system for final model construction, it was important to have chemicals in the training data set initially scored on a continuous scale to allow stepwise variation in the BP to identify an optimal cutoff. Furthermore, the MC4PC program gives greater weight to chemicals with marginal activities (20–29 activity units), recognizing them as having potential “weak positive” signals and allowing further refinement the process of identifying molecular features associated with activity. None of the programs were permitted to remove drugs from the training or validation data sets because they were statistical outliers. The same chemical structures in the training data sets were represented differently for the QSAR programs. MC4PC and MDL-QSAR programs used chemical structures represented as SMILES codes. In contrast, BioEpisteme and LPDM used chemical structures represented as MDL “.mol” files.

The four QSAR programs use distinctly different QSAR prediction paradigms, methods to access the QSAR model domain of applicability, and procedures to conduct leave-many-out (LMO) and leave-one-out (LOO) validation experiments. Technical information describing the operation of the four programs has been provided in considerable detail elsewhere (Matthews et al., 2008) to meet the requirements of the proposed OECD Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models (OECD, 2007).

2.3. Adjustment of low frequencies of active AEs

All of the individual specific and composite hepatobiliary and urinary tract AE endpoints had a relatively low percentage of actives (<20%; Ursem et al., 2009). A low percentage of actives, or a low active/inactive (A/I) ratio, can result in poor QSAR models with low sensitivity. This problem can be ameliorated, however, by increasing the percentage of actives by removing a fraction of the inactives while keeping the number of actives constant (Matthews et al., 2008). For example, half the inactive compounds can be distributed into two separate training sets, each containing the same active compounds. QSAR models can then be generated for the two training datasets with improved A/I ratios and the results of experiments performed with the two sets combined. The ratio of actives to inactives was not randomly adjusted; the inactive chemicals were first sorted by their SMILES codes to obtain a list with structurally similar chemicals adjacent. Then every second, third, fourth, etc. chemical was placed in subgroups in order to create two, three, four, etc., respectively, subsets of chemicals that had a similar set of chemical structures. The lower the frequency of actives in the control database, the higher the fold-adjustment that was required to achieve the same overall percentage of actives in the QSAR model. In this study this A/I adjustment procedure was used to improve the predictive performance of the MC4PC, BioEpisteme, and LPDM QSAR models. Since the optimal percentage of actives differed for these programs, the A/I ratio was adjusted to approximately 40% for MC4PC and 30% for BioEpisteme and LPDM models. For MDL-QSAR, adjusting the A/I ratio for data sets having a low percentage of actives had a detrimental effect on model performance and so this technique was not used for model optimization. The explanation for this phenomenon is likely to be related to MDL-QSAR's prediction paradigm which defines separate membership in class domains for both the active and inactive molecules. In this situation the ratio of actives and inactives is less important.

2.4. Identification of optimal BP for active and inactive drug AEs

The drug AEs for specific and composite hepatobiliary and urinary tract AE endpoints form a continuous distribution of PRR values over a 6-log range. We converted the PRR values into activity scores of 10–80 units, a scoring convention previously used with the MC4PC program (Matthews and Contrera, 1998), with inactive drugs being assigned 10–19 units, marginally active drugs assigned 20–29 units, and active chemicals assigned 30–80 units (Ursem et al., 2009). None of the specific and composite AE data sets had an obvious BP where the PRR-base activity unit value clearly separated active from inactive drugs. The lack of an obvious BP for AE data posed a technical problem for all four QSAR software programs because they rely on a binary (active and inactive) classification of toxicological response data. We, therefore, determined optimal BPs experimentally by systematically assessing the predictive performance of each AE QSAR model using incrementally different BP values to separate active and inactive drugs. Incremental BP values were increased by 2 activity units over the range of 12–40 units, i.e., 12, 14, 16 units, etc. and tested. This range was selected because nearly all drug AEs have activity unit values within this range. For each BP the predictive performance of an AE QSAR model was assessed in terms of the sensitivity and false positives, and a statistic was computed to compare the predictive performance of each set of adjacent BP values used in the experiment. The rate of change in the predictive performance (RCPP) was calculated as the change in the percent sensitivity divided by the change in percent false positives ($\Delta Se/\Delta FP$) per 1% change of the number of molecules ($\Delta \%Act$) in the training data set defined as positive [$RCPP = (\Delta Se/\Delta FP)/\Delta \%Act$].

In theory, all four QSAR programs could have been used to determine an optimal BP value for each AE endpoint. However, using all four would have required the preparation of 990 QSAR models for optimization experiments (15 models/composite endpoint \times 11 composite endpoints \times 6 experimental parameters), and the computer time necessary to perform these LMO validation experiments with all four programs would be formidable. Therefore BioEpisteme alone was selected for the optimization experiments because the analysis of a single 10% LMO experiment with a 1600 chemical data set takes only about 1 min in our hands. In contrast, a comparable LMO experiment using the same computer requires \sim 10–30 h for MC4PC, \sim 48 h for MDL-QSAR, and 1–6 h for LPDM. The BP identified in the BioEpisteme experiments was then confirmed in a second limited series of experiments using MC4PC.

2.5. Identification of the optimal method to classify AEs

We examined the impact of using different criteria to classify an AE response upon the predictive performance of the BioEpisteme QSAR program. The study examined the effect of different PRR methods employed by pharmacovigilance groups to detect serious drug-related AEs on QSAR model performance. The three PRR experimental parameters were: (1) varying the number of reports required for an AE to be called significant, (2) the use of a χ^2 statistic, and (3) inclusion of a Yates correction of the χ^2 statistics [Yates correction = (observed – expected – 0.5)²/expected]. In each experiment, a total of 15 different QSAR models were constructed using 2 activity unit increments over the range of 12–40 units. The composite kidney nephropathy endpoint was chosen for the experiments because preliminary studies showed good QSAR model predictive performance.

2.6. Hepatotoxicity external validation experiment

An external validation experiment was performed using a small set of drugs that had not been considered in the preparation of any

of the AE QSAR models. The data were obtained from a study that reviewed all of the safety withdrawals of prescription drugs from worldwide markets from 1960 to 1999 (Fung et al., 2001). Although the majority of the drugs listed in this study were already in our database (Ursem et al., 2009), a subset of 18 drugs showing serious hepatotoxicity effects were not present and were tested with all four QSAR programs. External validation is considered the most rigorous means for determining the predictive power and reliability of a QSAR model (Perkins et al., 2003; Golbraikh and Tropsha, 2002).

2.7. Statistical analyses

The predictive performance of the AE QSAR models was assessed through statistical comparisons using the method of Cooper et al. (1979), and the Pearson's χ^2 test, also known as the χ^2 goodness-of-fit test (see, for example, <http://math.hws.edu/javamath/ryan/ChiSquare.html>). In addition, we calculated a receiver operating characteristic intercept (ROC) statistic defined as sensitivity divided by 1-specificity (false positive), which was first used to characterize the detection of radio signals in the presence of noise in the 1940s (Lusted, 1971). Our criteria for the best possible AE models are: high specificity (>80%), high ROC intercept values (>2.0) with a low false positive rate (<20%). A poor model with no discriminating power has equal rates of true and false positives, resulting in a ROC value of 1 (Hanley, 1989; Provost and Fawcett, 2001; Benigni, 2005; Matthews et al., 2007a,b).

2.8. Availability of experimental data

The majority of the AE data compiled for this investigation were obtained from publicly available sources. SRS and AERS databases are available through the US Department of Commerce National Technical Information Service (<http://www.ntis.gov>). The AERS and clinical trial AE data (1992 to present) are also available in Elsevier's PharmaPendium database (<http://www.pharmapendium.com>). Additional AE data from the literature and drug labeling can be obtained using MicroMedex Integrated Index. The chemical structures of the pharmaceuticals are readily available through ChemIDplus (<http://chem.sis.nih.gov/chemidplus>). All of the AE QSAR models we constructed with these AE data are protected under licensing agreements and will be available to the scientific community through our CRADA partners, subject to appropriate approval.

3. Results

3.1. Identification of the optimal fraction of active chemicals in a MC4PC model

To determine the optimal fraction of active chemicals in a test set needed to obtain the best predictive performance, 33 MC4PC QSAR models were constructed to identify the optimal BPs for three composite hepatobiliary endpoint clusters (liver damage, liver enzyme increase, and jaundice) using AERS AE data. The predictive performance of the QSAR models was assessed when the percentage of active drugs was varied from 5% to 35% at 5% increments. The results of this experiment demonstrated that the BP selected affected the predictive performance of MC4PC QSAR models, and that BP values defining 20–25% of the chemicals as actives yielded the best AE QSAR models. Using a 20% BP, the MC4PC AE models had high specificity (88.8–93.2%), but relatively low sensitivity (SE, 23.0–31.1%). However, there was even less sensitivity in the QSAR models with BP values of 5% (SE = 8.7%), 10% (SE = 13.4%), 15% (SE = 14.1%), 25% (SE = 18.6%), 30% (SE = 15.2%), and 35% (SE = 15.0%). Based upon these data, a 20% BP value was considered

optimal and used in all MC4PC QSAR experiments employing a fixed BP.

3.2. Identification of the best source of AE data

134 MC4PC QSAR models were used to evaluate the effect of the source of the AE data (AERS or SRS) on the predictive performance of the QSAR models. Models were constructed from SRS AE data alone, AERS data alone, a composite of AERS and SRS data, and a composite of AERS and SRS data supplemented with AEs derived from clinical trials reported in the published literature. The models included 11 composite hepatobiliary and urinary tract AE endpoints and 11 specific hepatobiliary and urinary tract endpoints. The results showed that the source of the AE data had a small effect on the sensitivity of the models. Models using SRS data had slightly higher sensitivity than those using AERS data. Similarly, the AE data derived from a composite AERS and SRS supplemented with AE information from the literature resulted in a model that had slightly higher sensitivity than models derived from the AERS or SRS data sets alone. The most likely explanation of these results is that the larger the training data set for the AE model, the better the predictive performance of the model. The QSAR models derived from the AERS data alone had 1044–1056 chemicals compared to 1403–1412 chemicals in the SRS data alone models, and 1595–1608 chemicals in the composite QSAR models. In contrast, the results also demonstrated that the use of specific vs. composite endpoint data had no consistent effect on the predictive performance of QSAR models. In some cases the specific endpoint QSAR model had slightly higher sensitivity than the corresponding composite endpoints (e.g., interstitial nephritis vs. nephropathies). In other cases the specific endpoint had slightly lower sensitivity than the corresponding composite endpoint (e.g., jaundice and cholestasis vs. jaundice). Based upon these data, the composite database containing AERS, SRS, and literature data was used in all subsequent MC4PC QSAR experiments. A summary of all of these experiments revealed that individual MC4PC models had good performance in terms of high specificity ($90.5 \pm 5.2\%$), high coverage ($91.2 \pm 3.6\%$), low false positives (9.5%), and acceptable ROC intercept statistics (2.09 ± 0.52), but they had low sensitivity ($19.2 \pm 9.6\%$).

3.3. Identification of the optimal method to merge AERS and SRS AE data sets

Because the performance of models improved when the size of the training data set was increased, an experiment was conducted to determine the best method for merging the AERS and SRS AE data. This study used the same PRR method with 20% of the chemicals defined as actives for both data sets. Several hundred QSAR models were constructed and evaluated with BioEpisteme. The model performance was compared when the activity unit values for drugs in the AERS and SRS data sets were averaged, or when the higher activity unit value in either the AERS or SRS data set was used. The results of the several hundred BioEpisteme experiments revealed that the sensitivity of models containing the higher AERS or SRS activity unit value was consistently better than models using the average SRS and AERS value. Based upon these data, the higher AERS or SRS value was used in all subsequent QSAR experiments.

3.4. Effect of PRR method on performance of BioEpisteme QSAR models

We hypothesized that different methods for classification of significant drug AEs might also affect the performance of QSAR models, and one of these methods might provide superior model performance. Therefore, a series of experiments was performed

to study the effect of different PRR experimental parameters on BioEpisteme QSAR model performance.

The results of the LMO validation experiments for 180 BioEpisteme QSAR models are summarized in Table 1. The data are only presented for models with BPs of 30 and 32 activity units because these values were found to be optimal for both the BioEpisteme and MC4PC programs' nephropathy models based on their Cooper statistics. The results showed that two of the PRR method parameters had a very significant effect on the sensitivity of the BioEpisteme QSAR models. In the first series of experiments we examined the effect of varying the minimal number of reports required to define an AE as significant on the QSAR model performance. The results showed that the highest QSAR sensitivities were obtained when a minimum report requirement was set at 1–4 reports, with a 2 AE report minimum being optimal, having 19.7–24.4% sensitivity. In contrast, the sensitivities of models with 5 and 10 report minimum requirements were dramatically reduced (5.6–11.5%). These data also demonstrated that the use of the Yates correction for the χ^2 statistic consistently decreased the performance of the QSAR models made with a 3 report requirement. In contrast, these data showed that the addition of literature AE data for the composite kidney nephropathies endpoint had little or no effect on the performance of the nephropathy QSAR models. Taken together, these data showed that the PRR method that yielded the best QSAR models had a minimum requirement of 2 AE reports, a χ^2 of 4, but no Yates correction. These PRR parameters were used in all subsequent AE QSAR model experiments.

3.5. Optimal BP value for BioEpisteme hepatobiliary and urinary tract QSAR models

Using the PRR method identified above, a series of experiments was performed to estimate the optimal BP values for the 5 hepatobiliary and 6 urinary tract composite AE endpoints. These experiments were performed using the BioEpisteme program. A total of 165 QSAR models were constructed with 15 QSAR models for each of the 11 composite AE endpoints. The BP values were set at 1 or 2 activity unit increments from 12 to 40 units. The BP values of 10, the lowest possible value, and BP values of 50–80, the highest pos-

sible values, were not routinely used because optimal BP values were always observed within the range of 12–40. The optimal BP values were defined to be the value that showed a RCPP of the QSAR models between 2 consecutive BP values that was significantly higher (and positive) than those preceding.

The results of these experiments showed that the optimal BP values for the 5 hepatobiliary and 6 urinary tract composite AE endpoints were different for each endpoint (Table 2). The optimal BP values for the 5 hepatobiliary composite AE endpoints were 32 for liver disorders, 34 for jaundice and cholestasis, 28 for liver enzymes, 24 for gall bladder disorders, and 16 for bile duct disorders. Likewise, the optimal BP values for the 6 urinary tract composite AE endpoints were 30 for acute kidney disorders, 28 for kidney function tests, 26 for nephropathies, 26 for blood in urine, 20 for urolithiasis, and 31 for bladder disorders.

3.6. Optimal BP value for MC4PC hepatobiliary and urinary tract QSAR models

In an effort to generalize the methodology for constructing AE QSAR models, a series of experiments was performed to estimate the optimal BP values for the 5 hepatobiliary and 6 urinary tract composite AE endpoints using a second QSAR program. A total of 47 QSAR models were constructed for MC4PC including at least 3 QSAR models for each AE endpoint, and the BP values were set at intervals of 1 or 2 activity unit increments below and above the optimal BP value identified by BioEpisteme. The optimal BP values were evaluated in terms of the RCPP of the QSAR models.

The results of these experiments showed that the optimal BP values for the 5 hepatobiliary and 6 urinary tract composite AE endpoints were different for each endpoint (Table 3), and in most cases only slightly different from the optimal BPs identified for BioEpisteme QSAR models. The optimal MC4PC BP values for the 5 hepatobiliary composite AE endpoints were: 34 for liver disorders, 34 for jaundice and cholestasis, 36 for liver enzymes, 30 for gall bladder disorders, and 26 for bile duct disorders. Likewise, the study showed that the optimal BP values for the 6 urinary tract composite AE endpoints were 36 for acute kidney disorders, 28 for kidney function tests, 30 for nephropathies, 28 for blood in ur-

Table 1
Comparison of different PRR parameters on the predictive performance of BioEpisteme QSAR models.

Adverse effects		PRR parameters ^a				AE database ^b		Statistics	
Organ system	Composite endpoint	Minimum # reports	χ^2	Yates correct.	Used literat.	BP Activity	% Active	% Sensitivity	% False positives
Kidney	Nephropathies	1	No	No	No	32	9.9	17.2	3.9
					Yes		10.7	23.5	4.4
		2	4	No	No		10.5	20.4	5.3
					Yes		11.2	19.7	5.6
		3	4	No	No		8.8	14.3	4.3
					Yes		9.5	17.2	4.3
		3	4	Yes	No		8.5	11.9	4.5
		4	4	No	No		8.3	18.3	3.8
		5	4	No	No		6.5	11.5	3.6
		10	4	No	No	4.5	5.6	1.5	
		1	No	No	No	30	11.8	23.4	4.1
					Yes		12.7	23.4	5.0
		2	4	No	No		12.2	24.2	6.8
					Yes		12.9	24.4	6.9
		3	4	No	No		10.5	19.8	4.6
					Yes		11.2	19.7	5.2
		3	4	Yes	No		10.1	18.1	5.0
		4	4	No	No		9.7	20.8	4.6
		5	4	No	No		7.6	11.5	3.7
10	4	No	No	5.8	8.6	2.3			

Bold type indicates best results.

^a The minimum number of AE reports cut-off used, whether a χ^2 statistic was used, whether a Yates correction of the χ^2 statistic was used, and whether AEs from the literature data were used.

^b The BP used to separate active and inactive drugs and the % active drugs in the data set tested.

Table 2
Identification of the optimal BioEpisteme BP for hepatobiliary and urinary tract AE QSAR models.

Adverse effects		AE database ^a			Statistics				
Organ system	Composite endpoint	#Chem.	BP Activity	% Active	% Sensitivity	% False positives	ROC	RCPP	
<i>Hepatobiliary disorders</i>									
Liver	Liver disorders	1608	12	53.4	63.9	53.1	1.20		
			14	49.6	60.1	46.4	1.29	0.16	
			16	45.6	56.9	40.0	1.42	0.12	
			18	41.2	50.5	34.3	1.47	0.25	
			20	36.7	45.2	28.9	1.56	0.22	
			22	32.5	38.9	22.8	1.70	0.25	
			24	27.7	32.5	16.9	1.93	0.22	
			26	24.0	26.8	13.1	2.04	0.42	
			28	20.2	24.6	10.2	2.40	0.19	
			30	16.6	20.2	7.3	2.76	0.42	
							<i>Mean ± Stdev</i>	<i>0.25 ± 0.10</i>	
			31	15.7	20.6	6.8	3.02	-0.73	
			32	13.1	14.3	5.7	2.52	2.05	
			34 ^b	10.5	14.2	3.3	4.25	0.02	
			36	9.0	13.9	2.5	5.62	0.23	
			38	7.5	12.5	2.5	5.02	-46.57	
	40	5.5	7.9	1.4	5.66	2.18			
	Jaundice and cholestasis	1608	12	53.9	65.2	51.2	1.28		
			14	49.9	62.7	45.6	1.38	0.11	
			16	46.0	58.0	39.2	1.48	0.19	
			18	42.0	53.4	34.7	1.54	0.25	
			20	38.4	51.6	28.4	1.82	0.08	
			22	33.8	43.1	24.7	1.75	0.51	
			24	30.2	40.3	19.9	2.02	0.16	
			26	26.7	33.9	15.4	2.20	0.41	
			28	24.3	30.3	12.8	2.38	0.56	
			30	21.5	28.4	11.0	2.59	0.38	
			32	17.5	25.7	7.6	3.37	0.20	
							<i>Mean ± Stdev</i>	<i>0.28 ± 0.17</i>	
			34 ^b	14.8	22.7	6.4	3.56	0.90	
			36	12.7	21.1	4.6	4.54	0.44	
			38	10.9	17.7	4.0	4.44	2.87	
			40	8.3	13.4	3.2	4.20	2.12	
	Liver enzymes	1606	12	61.1	74.5	59.2	1.26		
			14	56.7	69.3	49.4	1.40	0.12	
			16	52.1	65.9	46.2	1.43	0.22	
			18	46.1	56.6	40.4	1.40	0.27	
			20	40.0	49.0	31.4	1.56	0.14	
			22	35.2	38.9	26.1	1.49	0.40	
			24	29.3	35.5	18.0	1.97	0.07	
26			24.5	29.3	12.9	2.28	0.25		
						<i>Mean ± Stdev</i>	<i>0.21 ± 0.11</i>		
28			21.7	24.7	9.9	2.50	0.55		
30			19.0	22.4	8.6	2.62	0.64		
31			17.0	20.9	7.2	2.92	0.53		
32			13.8	14.9	4.6	3.25	0.73		
34			11.0	17.1	4.1	4.12	-1.78		
36 ^b			8.5	7.4	2.1	3.48	1.92		
38			7.4	7.6	1.8	4.15	-0.68		
40	5.6	5.6	1.6	3.50	4.82				
Gall bladder	Gall bladder disorders	1056	12	44.8	48.8	32.1	1.52		
			14	41.1	47.6	28.3	1.68	0.09	
			16	36.6	43.9	22.3	1.97	0.13	
			18	31.6	35.7	19.3	1.85	0.55	
			20	27.7	33.0	16.0	2.06	0.21	
			22	23.9	27.0	12.3	2.19	0.43	
							<i>Mean ± Stdev</i>	<i>0.28 ± 0.20</i>	
			24	21.7	24.5	10.9	2.24	0.81	
			26	18.8	24.1	8.2	2.95	0.04	
			28	17.0	24.6	6.4	3.84	-0.14	
			30 ^b	15.2	16.9	6.3	2.70	30.57	
			32	12.5	16.7	4.0	4.16	0.04	
			34	8.8	8.6	2.6	3.31	1.55	
			36	7.1	9.3	1.9	4.81	-0.65	
			38	6.3	10.6	2.0	5.25	18.77	
			40	5.0	5.7	2.1	2.70	-50.26	
Bile duct	Bile duct disorders	1044	12	17.5	17.5	9.0	1.95		
			14	16.0	15.6	7.1	2.20	0.67	
							<i>Mean ± Stdev</i>	<i>NA</i>	
			16	15.0	14.0	5.9	2.38	1.35	
			18	13.8	11.1	5.0	2.22	2.68	
			20	11.8	17.1	4.4	3.92	-4.49	

(continued on next page)

Table 2 (continued)

Adverse effects		AE database ^a			Statistics					
Organ system	Composite endpoint	#Chem.	BP Activity	% Active	% Sensitivity	% False positives	ROC	RCPP		
Bile duct	Bile duct disorders	1044	22	10.6	11.7	4.0	2.95	12.27		
			24	9.6	12.0	2.7	4.53	-0.21		
			26 ^b	8.8	8.7	3.3	2.67	-7.06		
			28	7.7	11.3	2.3	4.91	-2.29		
			30	7.0	5.5	2.5	2.21	-45.29		
			32	6.2	9.2	1.9	4.76	-9.06		
			34	5.9	9.7	1.6	5.94	-5.05		
			36	5.3	10.9	1.6	6.73	-183.45		
			38	4.9	7.8	1.6	4.87	801.27		
			40	4.2	11.4	1.3	8.74	-16.93		
<i>Urinary tract disorders</i>										
Kidney	Kidney disorders	1595	12	62.6	72.3	64.5	1.12			
			14	57.7	65.7	57.7	1.14	0.20		
			16	52.1	58.2	50.2	1.16	0.18		
			18	44.8	49.0	40.7	1.20	0.13		
			20	39.5	43.8	30.4	1.44	0.10		
			22	34.2	38.7	24.7	1.56	0.17		
			24	29.4	30.8	19.8	1.55	0.34		
			26	24.1	25.9	15.0	1.73	0.19		
			28	20.1	20.7	11.7	1.77	0.39		
							<i>Mean ± Stdev</i>	<i>0.21 ± 0.10</i>		
			30	17.3	15.7	8.2	1.91	0.50		
			32	14.2	16.0	5.9	2.73	-0.04		
			34	11.4	12.8	3.8	3.34	0.56		
			36 ^b	9.0	11.3	3.4	3.33	1.41		
			38	7.0	8.3	2.2	3.82	1.19		
			40	5.5	7.0	1.8	3.90	2.40		
			Kidney function tests	1595	12	51.0	62.3	46.8	1.33	
					14	45.1	57.2	36.5	1.57	0.09
					16	39.2	51.2	29.6	1.73	0.14
					18	34.0	45.5	23.3	1.95	0.18
	20	29.3			40.3	19.6	2.05	0.30		
	22	25.2			36.8	16.5	2.22	0.28		
	24	21.3			32.1	13.2	2.44	0.35		
	26	18.4			27.5	9.4	2.93	0.42		
							<i>Mean ± Stdev</i>	<i>0.25 ± 0.12</i>		
	28 ^b	15.0			22.7	8.3	2.74	1.29		
	30	13.1			21.6	6.4	3.36	0.31		
	32	11.4			17.1	5.0	3.40	1.90		
	34	9.7	13.1	4.0	3.24	2.29				
	36	8.2	19.4	3.6	5.44	-8.92				
	38	6.3	17.2	1.9	9.13	0.70				
	40	4.5	15.5	1.1	13.83	1.26				
	Nephropathies	1595	12	38.4	43.3	32.1	1.35			
			14	34.5	44.0	25.2	1.75	-0.03		
			16	31.5	42.5	21.2	2.01	0.12		
			18	27.8	37.9	18.2	2.08	0.42		
			20	24.5	35.0	14.5	2.41	0.23		
			22	21.4	33.5	12.7	2.64	0.26		
			24	18.9	31.9	10.7	2.98	0.33		
							<i>Mean ± Stdev</i>	<i>0.23 ± 0.16</i>		
			26	16.7	28.2	8.9	3.16	0.94		
			28	14.5	22.5	7.0	3.22	1.35		
30 ^b			12.9	24.4	6.9	3.52	-19.22			
32			11.2	19.7	5.6	3.52	2.09			
34			9.7	17.0	3.7	4.60	0.90			
36			8.5	14.9	3.5	4.27	9.10			
38			6.9	11.9	2.9	4.11	3.19			
40			4.8	4.0	2.3	1.73	6.31			
Blood in urine			1595	12	43.6	51.4	37.3	1.38		
				14	40.2	47.2	34.6	1.36	0.45	
	16	34.6		40.0	25.2	1.59	0.14			
	18	28.8		32.4	19.4	1.67	0.23			
	20	24.3		26.4	15.1	1.75	0.31			
	22	19.4		24.9	10.4	2.40	0.07			
	24	15.5		19.9	7.2	2.76	0.40			
						<i>Mean ± Stdev</i>	<i>0.24 ± 0.15</i>			
	26	12.6		16.0	5.7	2.82	0.89			
	28 ^b	10.3		16.5	4.3	3.85	-0.14			
	30	8.7		12.3	3.0	4.16	1.94			
	32	6.2		6.1	1.9	3.26	2.29			
	34	4.4		4.4	0.9	4.73	1.01			
	36	3.3		2.0	0.7	3.02	7.84			
	38	2.5		2.6	0.5	5.69	-3.99			
	40	1.8		0.0	0.2	0.00	13.09			

Table 2 (continued)

Adverse effects		AE database ^a			Statistics				
Organ system	Composite endpoint	#Chem.	BP Activity	% Active	% Sensitivity	% False positives	ROC	RCPP	
Kidney	Urolithiasis	1595	12	26.2	29.2	18.7	1.56		
			14	22.3	26.8	15.4	1.74	0.18	
			16	18.5	21.8	10.6	2.05	0.28	
			18	14.8	16.1	6.4	2.51	0.36	
							<i>Mean ± Stdev</i>		0.27 ± 0.09
			20	12.5	9.6	4.7	2.04	1.61	
			22	10.6	11.2	3.7	3.07	−0.89	
			24 ^b	9.1	15.9	2.6	6.20	−2.79	
			26	7.6	14.1	1.8	7.64	1.67	
			28	6.8	18.5	2.0	9.17	30.47	
			30	5.6	10.0	1.3	7.52	10.94	
			32	3.8	9.8	1.1	9.37	0.31	
			34	2.6	11.9	0.2	62.63	−2.01	
			36	2.1	3.0	0.3	11.65	−224.57	
			Bladder	Bladder disorders	1595	12	51.9	61.5	47.4
14	49.0	58.2				44.0	1.32	0.33	
16	45.4	56.7				38.1	1.49	0.07	
18	41.5	50.3				35.2	1.43	0.57	
20	37.4	46.5				31.0	1.50	0.22	
22	33.4	41.3				26.7	1.55	0.30	
24	30.2	40.8				22.4	1.82	0.04	
26	25.8	31.4				17.0	1.85	0.40	
28	22.4	27.8				13.0	2.13	0.27	
30	19.7	26.2				9.9	2.66	0.19	
							<i>Mean ± Stdev</i>		0.26 ± 0.16
31	17.9	22.9				9.0	2.53	2.19	
32 ^b	15.0	15.1				8.0	1.88	2.67	
34	12.2	12.4				5.4	2.30	0.36	
36	10.6	13.1				3.6	3.66	−0.25	
38	8.8	14.3	2.8	5.05	−0.87				
40	7.1	9.7	1.8	5.53	2.52				

Bold type indicates best BP results obtained for the BioEpisteme QSAR models.

^a The number of chemicals in the model, the BP used to separate active and inactive drugs, and the % active drugs in the data set tested.

^b Optimal BP found for MC4PC (Table 3).

ine, 24 for urolithiasis, and 32 for bladder disorders. These data also showed that the optimal BP value for MC4PC (Table 3) was 0–10 units higher than the value for BioEpisteme (Table 2), being an average of 3.72 units higher. However, using the optimal MC4PC BP values with BioEpisteme still resulted in good QSAR models.

3.7. Comparison of performance of BioEpisteme, MC4PC, MDL-QSAR and Leadscope AE QSAR models

A series of experiments was performed to generalize the methodology for constructing AE QSAR models and to determine whether different QSAR program paradigms could be used to predict serious drug AEs. The experiments were conducted using two programs that employ molecular fragment prediction paradigms, MC4PC and LPDM, and two programs using molecular descriptor prediction paradigms, BioEpisteme and MDL-QSAR. All of these experiments were conducted using the same training data sets that employed the optimal BP values identified by MC4PC for the 5 hepatobiliary and 6 urinary tract composite AE endpoints (Table 3). The results for the MDL-QSAR program were obtained after the QSAR models had been optimized for performance using the smoothing parameter, r (Matthews et al., 2008). The results for the BioEpisteme and Leadscope programs are presented for two types of AE QSAR models in which the A/I ratio was either unadjusted or adjusted to optimize predictive performance. The BioEpisteme results for the unadjusted ($1\times$) A/I QSAR models at the MC4PC optimal BPs are presented in Table 2. The predictive performance of all four programs was optimized so that the QSAR models achieved the highest possible sensitivity when specificity was at least 80% (20% false positives).

The results of these experiments are summarized in Table 4. They demonstrate that AE QSAR models could be constructed for

all 11 composite AE endpoints using three QSAR programs (MC4PC, BioEpisteme and LPDM), and 9 of 11 endpoints using MDL-QSAR. MDL-QSAR models could not be generated for acute kidney disorders and bile duct disorders. For all four programs and all 11 composite AE endpoints, the average specificity, coverage, and ROC values were $86.5 \pm 1.7\%$; $92.4 \pm 2.6\%$; and 3.32 ± 0.39 , respectively. In contrast, the average sensitivity of the QSAR models for all four programs was $39.3 \pm 7.2\%$ for the 11 composite modules. These data also showed that the sensitivity of the QSAR models was notably different for individual AE endpoints. For example, relatively high sensitivity models could be constructed for kidney nephropathies using all four QSAR programs ($52.2 \pm 7.4\%$). In contrast, three of the QSAR programs had difficulty predicting bile duct disorders and the overall sensitivity of this composite endpoint was $30.8 \pm 9.8\%$. Taken together, these data demonstrated that each of the QSAR paradigms could be used to predict hepatobiliary and urinary tract AEs, but no one QSAR program had an overall superior predictive performance for all 11 composite endpoints compared to the other programs.

The results of these experiments also demonstrated that the predictive performance of certain QSAR programs was clearly influenced by the A/I ratio in the training data sets of the AE models. For example, the BioEpisteme models for kidney disorders had a sensitivity of 11.3% for an unadjusted A/I ratio, but 27.7% with a 3-fold increase in the A/I ratio. Comparable increases in sensitivity were observed for all of the other composite AE endpoints. Similarly, the LPDM models for kidney disorders had a sensitivity of 16.0% for an unadjusted A/I ratio, but 31.1% with a 3-fold increase in the A/I ratio. The predictive performance of the MC4PC QSAR models was also influenced by A/I ratio. Preliminary experiments using a fixed percentage of drugs with significant AEs (Ursem et al., 2009) for several of the composite AE endpoints produced

Table 3
Optimization of the predictive performance of MC4PC hepatobiliary and urinary tract QSAR models.

Adverse effects		AE database ^a			MC4PC ^b		Statistics			
Organ system	Composite endpoint	# Chem.	BP activity	% Actives	Fold adjust.	% Active	% Specificity	% Sensitivity	% Coverage	ROC
<i>Hepatobiliary disorders</i>										
Liver	Liver disorders	1608	32	13.1	4×	37.5	89.9	31.2	88.7	3.10
			34	10.5	5×	37.0	85.8	41.1	89.6	2.90
			36	9.0	6×	37.1	86.4	38.0	88.9	2.80
	Jaundice and cholestasis	1608	32	17.5	3×	38.8	93.3	25.2	94.2	3.78
			34	14.8	4×	41.0	87.8	37.8	90.4	3.09
			36	12.7	4×	36.7	91.8	27.6	90.6	3.37
	Liver enzymes	1606	30	19.0	3×	41.3	93.7	17.4	94.5	2.75
			31	17.0	3×	38.0	93.0	17.3	93.8	2.48
			32	13.8	4×	39.1	89.1	24.9	90.8	2.28
			34	11.0	5×	38.1	87.2	19.1	84.8	1.49
36			8.5	6×	35.7	84.5	33.7	87.4	2.17	
Gall bladder	Gall bladder disorders	1056	38	7.4	6×	32.4	87.3	32.0	85.9	2.52
			24	21.7	2.5×	40.8	82.8	27.6	89.1	1.60
			26	18.8	3×	41.0	91.6	21.0	91.5	2.50
			28	17.0	3×	37.9	92.3	19.2	91.1	2.48
			30	15.2	4×	41.7	86.3	30.7	86.2	2.24
	Bile duct disorders	1044	32	12.5	4×	36.4	90.8	27.2	85.9	2.94
			18	13.8	4×	39.0	89.9	27.1	89.5	2.68
			20	11.8	4×	34.5	89.7	19.6	84.2	1.90
			22	10.6	5×	37.3	88.2	22.7	82.6	1.93
			24	9.6	5×	34.6	88.4	21.1	83.7	1.83
26	8.8	6×	36.7	81.3	41.8	79.8	2.24			
28	7.7	6×	33.2	87.0	26.7	79.0	2.04			
<i>Urinary tract disorders</i>										
Kidney	Kidney disorders	1595	32	14.2	4×	39.9	89.9	22.4	90.8	2.22
			34	11.4	5×	39.1	86.1	25.1	90.3	1.81
			36	9.0	5×	37.3	80.4	42.6	87.5	2.18
			38	7.0	7×	34.4	82.6	33.2	87.0	1.91
			26	18.4	3×	40.3	91.8	29.1	94.9	3.53
	Kidney function tests	1595	28	15.0	4×	41.1	87.7	43.3	91.6	3.57
			28	14.0	4×	39.5	89.0	38.8	90.9	3.54
			30	13.1	4×	37.6	89.4	39.8	90.9	3.77
			28	14.6	4×	40.5	89.7	38.6	91.1	3.74
			30	12.9	5×	42.6	81.5	49.5	91.4	2.68
	Nephropathies	1595	32	11.2	5×	38.7	86.0	42.3	91.5	3.03
			26	12.6	4×	34.6	90.8	26.9	89.7	2.92
			28	10.3	6×	40.9	80.7	55.2	87.1	2.86
			30	8.7	6×	36.4	87.2	40.7	85.8	3.19
			32a	6.2	4×	34.6	91.5	28.8	81.9	3.41
	Blood in urine	1595	32b	6.2	4×	34.6	86.7	33.9	81.8	2.54
			22	11.2	5×	39.1	89.6	13.5	89.7	1.30
			24	9.1	6×	37.5	84.2	31.3	87.0	1.98
			26	7.6	7×	36.5	82.6	18.2	85.7	1.05
			30	8.7	6×	36.4	87.2	40.7	85.8	3.19
Bladder	Bladder Disorders	1595	32a	6.2	4×	34.6	91.5	28.8	81.9	3.41
			32b	6.2	4×	34.6	86.7	33.9	81.8	2.54
			22	11.2	5×	39.1	89.6	13.5	89.7	1.30
			24	9.1	6×	37.5	84.2	31.3	87.0	1.98
			26	7.6	7×	36.5	82.6	18.2	85.7	1.05
	Bladder Disorders	1595	30	19.7	3×	42.4	92.1	20.4	94.0	2.59
			31	17.9	3×	39.5	93.9	18.2	93.9	2.96
			32	15.1	4×	41.5	87.6	36.4	90.1	2.93
			34	12.2	5×	41.1	81.2	37.6	90.1	2.00

Bold type indicates best results.

^a The number of chemicals in the model, the % active drugs in the data set tested, and the BP used to separate active and inactive drugs.

^b MC4PC QSAR models had an adjustment of the A/I ratio and a higher % actives than the original AE data set.

QSAR models with good specificity, but low sensitivity of <15%. In contrast, adjustment of the A/I ratio did not improve the performance of MDL-QSAR models using a non-parametric discriminate analysis prediction paradigm.

3.8. Are the AE predictions of four QSAR programs confirmatory or complementary?

An additional analysis of the AE QSAR model prediction data described above was performed to determine whether the four QSAR programs were providing confirmatory or complementary predictions of AEs. If the QSAR programs predict nearly all of the same drugs to be either active or inactive, they would be evaluated as confirmatory. In contrast, if the QSAR programs predict different drugs to be either active or inactive, they could be considered complementary. If one or more of the QSAR programs were shown to

be complementary, then the predictions of these programs could be combined to possibly improve the overall performance of models to predict AEs. The investigation was performed with the six possible pairings of the four QSAR programs with the 11 composite AE endpoints. The confirmatory vs. complementary evaluations were made based upon the overall sensitivity and ROC statistics.

Because confirmatory programs predict the same chemicals to be active or inactive, they have a high sensitivity by predicting the same chemicals active (e.g., >80%), a high specificity by predicting the same chemicals inactive (e.g., >80%), and high ROC values (e.g., >10.0) which are calculated by dividing the sensitivity by one minus the specificity. In sharp contrast, complementary programs which predict different portions of the chemicals to be active or inactive would be expected to have a moderate sensitivity (e.g., 50%) when the two programs correctly predict about half of the known actives. Assuming the QSAR models have their recom-

Table 4

Summary of the best predictive performance of hepatobiliary and urinary tract QSAR models for four QSAR programs.

Adverse effects		AE database ^a			QSAR program	QSAR model ^b		Statistics				
Organ system	Composite endpoint	#Chem.	BP activity	% Actives		Fold adjust.	% Active	% Specificity	% Sensitivity	% Coverage	ROC	
Liver	Liver disorders	1608	34	10.5	MC4PC	5×	37.0	85.8	41.1	89.6	2.90	
					MDL-QSAR	1×	10.5	85.3	37.7	93.7	2.57	
					BioEpisteme	3×	26.1	86.2	28.8	NA ^d	2.09	
					LPDM	3×	26.0	91.0	49.1	97.1	5.44	
					<i>Mean ± Stdev^c</i>			87.1 ± 2.6	39.2 ± 8.4	93.5 ± 3.8	3.25 ± 1.50	
					BioEpisteme	1×	10.5	96.7	14.2	NA	4.30	
	Jaundice and cholestasis	1608	34	14.8	LPDM	1×	10.5	97.8	26.6	94.7	11.99	
					MC4PC	4×	41.0	87.8	37.8	90.4	3.09	
					MDL-QSAR	1×	14.8	85.5	43.6	95.1	3.01	
					BioEpisteme	2×	25.5	84.1	37.6	NA	2.41	
					LPDM	3×	34.2	88.9	57.6	98.8	5.18	
					<i>Mean ± Stdev</i>			86.6 ± 2.2	44.2 ± 9.4	94.8 ± 4.2	3.42 ± 1.21	
	Liver enzymes	1606	36	8.5	BioEpisteme	1×	14.8	93.6	22.7	NA	3.56	
					LPDM	1×	14.8	96.5	33.2	95.6	9.75	
					MC4PC	6×	35.7	84.5	33.7	87.4	2.17	
					MDL-QSAR	1×	8.5	84.0	28.9	97.3	1.81	
					BioEpisteme	2×	15.6	93.5	17.3	NA	2.48	
					LPDM	3×	19.6	93.9	34.6	95.7	5.69	
<i>Mean ± Stdev</i>			89.0 ± 5.5	28.6 ± 8.0	93.5 ± 5.3	3.04 ± 1.79						
Gall Bladder	Gall bladder disorders	1056	30	15.2	BioEpisteme	1×	8.5	97.9	7.4	NA	3.48	
					LPDM	1×	8.5	99.1	9.6	90.8	10.76	
					MC4PC	4×	41.7	86.3	30.7	86.2	2.24	
					MDL-QSAR	1×	15.2	85.1	36.6	94.1	2.43	
					BioEpisteme	2×	26.3	85.7	35.3	NA	2.41	
					LPDM	2×	26.3	90.5	46.9	98.5	4.94	
	<i>Mean ± Stdev</i>			86.9 ± 2.4	37.4 ± 6.8	92.9 ± 6.2	3.01 ± 1.29					
	Bile Duct	Bile duct disorders	1044	26	8.8	BioEpisteme	1×	15.2	93.7	16.9	NA	2.70
						LPDM	1×	15.2	96.3	21.9	95.1	5.93
						MC4PC	6×	36.7	81.3	41.8	79.8	2.24
						MDL-QSAR	1×	8.8	A model could not be constructed			
						BioEpisteme	3×	22.4	88.0	27.9	NA	2.36
LPDM						2×	16.2	96.8	22.8	91.5	7.24	
<i>Mean ± Stdev</i>			88.7 ± 7.8	30.8 ± 9.8	85.7 ± 8.3	3.95 ± 2.85						
Kidney	Kidney disorders	1595	36	9.0	BioEpisteme	1×	8.8	96.7	8.7	NA	2.67	
					LPDM	1×	8.8	98.6	20.7	95.6	15.11	
					MC4PC	5×	37.3	80.4	42.6	87.5	2.18	
					MDL-QSAR	1×	9.0	A model could not be constructed				
					BioEpisteme	3×	16.6	86.3	27.7	NA	2.02	
					LPDM	3×	23.0	95.4	31.3	95.2	6.81	
	<i>Mean ± Stdev^c</i>			87.4 ± 7.6	33.9 ± 7.8	91.4 ± 5.4	3.67 ± 2.72					
	Kidney function tests	1595	28	15.0	BioEpisteme	1×	9.0	96.6	11.3	NA	3.33	
					LPDM	1×	9.0	99.1	16.0	96.7	17.98	
					MC4PC	4×	41.1	87.7	43.3	91.6	3.57	
					MDL-QSAR	1×	15.0	89.0	36.2	88.3	3.29	
					BioEpisteme	2×	26.1	83.2	41.1	NA	2.46	
					LPDM	3×	34.6	88.5	55.9	98.0	4.84	
	<i>Mean ± Stdev</i>			87.1 ± 2.7	44.1 ± 8.4	90.0 ± 2.3	3.54 ± 0.99					
	Nephropathies	1595	30	12.9	BioEpisteme	1×	15.0	91.7	22.7	NA	2.74	
					LPDM	1×	15.0	94.7	40.8	93.5	7.76	
					MC4PC	5×	42.6	81.5	49.5	91.4	2.68	
					MDL-QSAR	1×	12.9	88.0	47.0	96.6	3.91	
BioEpisteme					3×	30.8	80.0	49.0	NA	2.48		
LPDM					3×	30.8	89.0	63.1	92.5	5.76		
<i>Mean ± Stdev</i>			84.6 ± 4.5	52.2 ± 7.4	93.5 ± 2.7	3.71 ± 1.51						
Blood in urine	1595	28	10.3	BioEpisteme	1×	12.9	93.1	24.4	NA	3.52		
				LPDM	1×	12.9	95.9	34.0	97.2	8.20		
				MC4PC	6×	40.8	80.7	55.2	87.1	2.86		
				MDL-QSAR	1×	10.3	78.7	43.0	97.9	2.02		
				BioEpisteme	3×	25.7	85.2	41.5	NA	2.83		
				LPDM	3×	25.7	91.0	49.1	96.7	5.44		
<i>Mean ± Stdev</i>			83.9 ± 5.5	47.2 ± 6.3	93.9 ± 5.9	3.29 ± 1.49						
Urolithiasis	1595	24	9.1	BioEpisteme	1×	10.3	95.7	16.5	NA	3.90		
				LPDM	1×	10.3	97.8	25.5	87.1	11.46		
				MC4PC	6×	37.5	84.2	31.3	87.0	1.98		
				MDL-QSAR	1×	9.1	83.0	38.3	95.5	2.26		
				BioEpisteme	3×	22.8	88.7	30.1	NA	2.72		
				LPDM	3×	23.1	91.5	41.1	94.9	4.95		
<i>Mean ± Stdev</i>			86.9 ± 4.0	35.2 ± 5.3	92.5 ± 4.7	2.98 ± 1.35						
BioEpisteme	1×	9.1	97.4	15.9	NA	6.20						
LPDM	1×	9.1	98.7	8.3	88.1	8.31						

(continued on next page)

Table 4 (continued)

Adverse effects		AE database ^a			QSAR program	QSAR model ^b		Statistics			
Organ system	Composite endpoint	#Chem.	BP activity	% Actives		Fold adjust.	% Active	% Specificity	% Sensitivity	% Coverage	ROC
Bladder	Bladder disorders	1595	32	15.1	MC4PC	4×	41.5	87.6	36.4	90.1	2.93
					MDL-QSAR	1×	15.1	78.2	34.7	96.7	1.59
					BioEpisteme	2×	26.1	82.3	35.9	NA	2.03
					LPDM	3×	34.7	87.0	51.5	98.0	4.00
					<i>Mean ± Stdev</i>		83.8 ± 4.4	39.6 ± 7.9	94.9 ± 4.2	2.64 ± 1.07	
					BioEpisteme	1×	15.1	92.0	15.1	NA	1.88
					LPDM	1×	15.1	98.7	10.5	92.5	8.31
4 program overall Mean ± Stdev								86.5 ± 1.7	39.3 ± 7.2	92.4 ± 2.6	3.32 ± 0.39

^a The number of chemicals in the model, the % active drugs in the data set tested, and the BP used to separate active and inactive drugs.

^b A portion of the QSAR models had an adjustment of the A/I ratio and a higher % actives than the original AE data set.

^c The mean and standard deviation are presented for the best QSAR models for the four QSAR programs.

^d The coverage of the BioEpisteme model could not be determined precisely, but it was estimated to be >95%.

mended specificity of 85%, the ROC values of complementary models would be expected to be in the range of 2–5.

The results of this study demonstrated that the AE QSAR models for the four QSAR programs were all complementary (Table 5). Each pair of QSAR programs had comparable specificities, sensitivities, ROC values, and χ^2 values, and none of the program pairs had markedly higher statistical parameters compared to any of the other pairs of programs. The average specificity, sen-

sitivity, ROC value, and χ^2 value for the 6 pairs of QSAR programs were $87.8 \pm 2.2\%$, $45.4 \pm 5.4\%$, $4.00 \pm 1.01\%$, and $158.0 \pm 47.8\%$, respectively. Although not statistically significant, there was an elevation in both the ROC and χ^2 values for the MC4PC and LPDM pair and less so for the LPDM and BioEpisteme pair, suggesting that the prediction paradigms of these program pairs may have more in common than the remaining QSAR programs.

Table 5

Summary of the ability of QSAR Program #1 to predict QSAR Program #2.

QSAR Program		Statistics			
No. 1	No. 2	% Specificity	% Sensitivity	ROC	χ^2
MC4PC	MDL-QSAR	86.1	43.8	3.31	110.6
MC4PC	BioEpisteme	87.7	43.1	3.56	134.0
MC4PC	LPDM	92.1	41.4	5.84	196.1
BioEpisteme	MDL-QSAR	87.4	43.1	3.40	139.8
LPDM	MDL-QSAR	85.9	45.0	3.51	131.6
LPDM	BioEpisteme	87.8	56.1	4.66	236.1
<i>Mean ± Stdev</i>		87.8 ± 2.3	45.4 ± 5.4	4.00 ± 1.01	158.0 ± 47.8

Table 6

Results of using two QSAR programs to predict 11 AE endpoints.

QSAR Program		Statistics				
No. 1	No. 2	% Specificity	% Sensitivity	ROC	χ^2	
MC4PC	LPDM	82.0 ± 1.5	52.8 ± 12.1	2.92 ± 0.52	124.6 ± 78.4	
MC4PC	BioEpisteme	76.8 ± 3.0	56.0 ± 9.2	2.41 ± 0.20	89.7 ± 44.0	
MC4PC	MDL-QSAR	76.3 ± 3.7	51.8 ± 6.8	2.23 ± 0.45	73.9 ± 47.9	
LPDM	BioEpisteme	81.5 ± 4.4	59.2 ± 12.5	3.23 ± 0.24	150.6 ± 62.1	
LPDM	MDL-QSAR	78.8 ± 3.4	60.1 ± 7.5	2.90 ± 0.60	141.5 ± 79.0	
BioEpisteme	MDL-QSAR	75.2 ± 3.4	57.5 ± 9.4	2.32 ± 0.39	91.5 ± 48.6	
<i>Mean ± Stdev</i>		78.4 ± 3.2	56.2 ± 9.6	2.67 ± 0.39	112.0 ± 60.0	
Using one program		Mean ± Stdev	86.5 ± 1.7	39.3 ± 7.2	3.32 ± 0.39	NA

Table 7

Comparison of the predictive performance of AE QSAR models using one or more QSAR prediction paradigms.

No. of QSAR programs	No. of positive predictions ^a	Ave. Drug AE Score ^b	Statistics			
			% Specificity	% Sensitivity	ROC	χ^2
4	4	41.5 ± 3.3	98.5 ± 0.7	13.0 ± 6.5	9.21 ± 4.23	82.5 ± 54.3
4	3 or more	43.2 ± 3.4	95.7 ± 2.1	27.9 ± 12.3	7.25 ± 2.07	143.0 ± 79.6
4	2 or more	42.5 ± 2.6	89.0 ± 3.8	46.3 ± 12.2	4.41 ± 0.86	157.7 ± 67.3
4	1 or more	42.1 ± 2.5	68.6 ± 3.8	67.9 ± 10.2	2.16 ± 0.21	97.9 ± 52.6
2	1	41.6 ± 2.3	78.4 ± 3.2	56.2 ± 9.6	2.67 ± 0.39	NA
1	1	41.6 ± 2.3	86.5 ± 1.7	39.3 ± 7.2	3.32 ± 0.39	NA

Bold type indicates best results.

^a The number of positive predictions required to call a drug active.

^b The average score of the active drugs.

Table 8A

Summary of the results of the hepatotoxicity external validation experiment.

Pharmaceutical generic name	Market-terminating Adverse Effect	MC4PC expert call	MDL-QSAR expert call	BioEpisteme expert call	LPDM expert call	Positive QSAR models ^a	Positive consensus programs ^b
bendazac	Hepatotoxicity	—	3+	6+	4+	13+	3+
benzarone	Hepatitis	1+	—	—	—	1+	1+
benziodarone	Jaundice	1+	—	1+	5+	7+	3+
cyclofenil	Hepatotoxicity	—	—	4+	—	4+	1+
ebrotidine	Hepatotoxicity	3+	—	5+	1+	9+	3+
exifone	Hepatotoxicity	3+	1+	2+	3+	9+	4+
fipexide	Hepatotoxicity	1+	2+	—	—	3+	2+
ibufenac	Hepatotoxicity; jaundice	2+	1+	2+	—	5+	3+
isaxoninephosphate	Hepatotoxicity	2+	1+	1+	4+	8+	4+
mebanazine	Hepatotoxicity	—	—	—	Not predicted	0+	0+
nialamide	Hepatotoxicity	3+	3+	11+	—	17+	3+
nitrefazole	Hepatotoxicity	—	3+	5+	—	8+	2+
phenoxypropazine	Hepatotoxicity	—	3+	—	—	3+	1+
sulfacarbamide	Hepatic reactions	—	1+	—	—	1+	1+
suloctidyl	Hepatotoxicity	—	—	—	—	0+	0+
triacetyldiphenolisatin	Hepatotoxicity	—	—	1+	8+	9+	2+
xenazoicacid	Hepatotoxicity	3+	3+	3+	—	9+	3+
zimeldine	Hepatotoxicity	1+	1+	—	—	2+	2+

^a The number of QSAR models with positive predictions.^b The number of QSAR programs that predicted the test chemicals to be positive.**Table 8B**

Summary of the statistics of the hepatotoxicity external validation experiment.

# QSAR programs	Positive predictions	
	Average # Active ^a	Average % Sensitivity ^b
any 1	10.5	52.8
any 2	13.3	74.1
any 3	15.3	85.2
all 4	16.0	88.9

^a The average number of drugs predicted active was calculated using any 1, 2, 3, or all 4 QSAR programs.^b The percent (%) sensitivity of using 1 to 4 QSAR programs was calculated using the method of Cooper et al. (1979).

3.9. Results of consensus prediction experiments using two QSAR paradigms

Since all four QSAR programs were found to be complementary, we hypothesized that the overall predictive performance of the AE QSAR might be improved if the predictions from two programs were combined. In this situation a positive prediction from either program would be counted. In particular, the sensitivity of predicting AEs might be enhanced by using complementary programs because they might be detecting different active drugs with significant findings. This investigation was performed using the

AE prediction results described in Section 3.7 and the six pairs of QSAR programs described in Section 3.8. The results of this study demonstrated that the predictive performance of the six pairs of programs were remarkably similar and the average specificity, sensitivity, ROC value, and χ^2 values were $78.4 \pm 3.2\%$, $56.2 \pm 9.6\%$, $2.67 \pm 0.39\%$, and $112.0 \pm 60.0\%$, respectively (Table 6). No pair of programs had substantially better performance compared to the other pairs of programs. The results also demonstrated that the predictive performance of a single AE QSAR model was substantially improved by combining the predictions of two QSAR programs. The average sensitivity of using two programs vs. one program alone was improved nearly 17% (56.2% vs. 39.3%). In contrast, the overall specificity of using two QSAR programs was only diminished by about 8% (78.4% vs. 86.5%). Likewise, the overall ROC value of using two QSAR programs was only slightly reduced (2.67 vs. 3.32).

3.10. Can consensus QSAR program predictions provide increased confidence?

Since all four QSAR programs were shown to be complementary and predictions from two QSAR programs had improved predictive performance, we hypothesized that the relative confidence in AE positive predictions might also be improved by combining confirmatory predictions from two or more programs. For example, if a

Table 9

Summary of the MC4PC QSAR model numbers for the 11 hepatobiliary and urinary tract composite AE endpoints.

Adverse effects		MC4PC AE models ^a		
Organ system	Composite endpoint	Fold adjust.	Control model	Model set
Liver	Liver disorders	5×	A10	A11,A12,A13,A14,A15
	Jaundice and cholestasis	4×	A25	A26,A27,A28,A29
	Liver enzymes	6×	A17	A18,A19,A20,A21,A22,A23
Gall bladder	Gall bladder disorders	4×	A31	A32,A33,A34,A35
Bile duct	Bile duct disorders	6×	A37	A38,A39,A40,A41,A42,A43
Kidney	Kidney disorders	5×	A10	A11,A12,A13,A14,A15
	Nephropathies	5×	A17	A18,A19,A20,A21,A22
	Kidney function tests	4×	A24	A25,A26,A27,A28
	Blood in urine	6×	A30	A31,A32,A33,A34,A35,A36
Bladder	Urolithiasis	6×	A38	A39,A40,A41,A42,A43,A44
	Bladder disorders	4×	A46	A47,A48,A49,A50

^a A portion of the QSAR models had an adjustment of the A/I ratio and a higher % actives than the original AE data set.

drug was predicted active at a given AE endpoint by two different QSAR program paradigms, and each program had been validated and had acceptable performance statistics for this endpoint, the drug with two positive predictions would have an enhanced probability of causing this AE. Similarly, if a drug was predicted active with three or four QSAR program paradigms, this drug with three or four positive predictions would have an even higher probability of truly being the cause of this AE. This investigation was also performed using the AE prediction results described in Section 3.7 and the six pairs of QSAR programs described in Section 3.8.

The results of this study confirmed this hypothesis and showed that the predictive performance of consensus predictions for the AE QSAR models was substantially improved by combining two or more QSAR programs compared to using one program alone (Table 7). The average specificity obtained when using consensus predictions from 2, 3, or 4 programs vs. 1 program alone was increased from 86.5% to 89.0% (2 programs), 95.7% (3 programs), and 98.5% (4 programs). Similarly, the average ROC value obtained when using consensus predictions from 2, 3, or 4 programs vs. 1 program alone was improved from 3.32 to 4.41 (2 programs), 7.25 (3 programs), and 9.21 (4 programs). Although the increasing specificity and ROC values is desirable, these data also showed the average sensitivity obtained when using consensus predictions requiring agreement among 2, 3, or 4 programs vs. calling a chemical positive if found positive in either of 2 programs was diminished from 56.2% to 46.3% (2 programs), 27.9% (3 programs), and 13.0% (4 programs). In contrast, the χ^2 values for the combining of 2 or more QSAR programs compared to 1 program alone were relatively similar, but the 2 program experimental condition had the highest value of 157.7. This indicated that the program set had remarkably significant predictions of AEs and in our opinion the 2 program combination was considered to be optimal.

Since unanimous consensus positive predictions obtained using 2, 3, or 4 QSAR programs had progressively increased specificity and ROC values, it was hypothesized that consensus predictions might also identify drugs that had been assigned a higher WOE for AEs. However, the results of this study showed that the subset of active drugs having consensus AE predictions did not have higher activities compared to active drugs that were not predicted positive (column 3, Table 7).

3.11. External validation study of pharmaceutical hepatotoxicity

An external validation experiment was performed using a set of 18 drugs known to induce serious hepatotoxicity that have been removed from worldwide markets, but were not part of the training set used for this study. The QSAR predictions were obtained using each of the four QSAR programs. The results showed that 16 out of 18 (88.9%) of the drugs were predicted to be hepatotoxic by at least 1 of the 4 programs (Table 8A). Furthermore, 12 out of 18 (66.7%), 8 out of 18 (44.4%), and 2 out of 18 (11.1%) of the hepatotoxic drugs were detected by consensus predictions in at least 2, 3, and all 4 programs, respectively.

3.12. Recommended hepatobiliary and urinary tract human AE models

Table 9 presents the numbers of the QSAR models to be used with the MC4PC software program to predict possible human adverse hepatobiliary and urinary tract events. For example, to obtain a prediction for liver disorders, the set of six models (A10, A11, A12, A13, A14, and A15) would be run together and evaluated using our ICSAS expert rules (Matthews et al., 2008). The models are designed to predict the 11 hepatobiliary and urinary tract composite AE endpoints described in this investigation (Ursem et al., 2009). To obtain the best results we recommended the use of all 11 of the AE QSAR models and at least two different QSAR pro-

grams to provide a WOE prediction of test chemicals for potential hepatobiliary and urinary tract disorders. For example, if a test chemical is predicted active by one QSAR program and one AE QSAR model for a single organ system, this is sufficient evidence to evaluate the chemical as a weak positive. However, if the test chemical is predicted active in more than one AE QSAR model, or more than one QSAR program, this is sufficient evidence to give it an even higher level of concern, a strong positive. The model numbers for the remaining three programs will be posted and updated at our website (http://www.fda.gov/cder/Offices/OPS_IO/default.htm) when they become available to the scientific community through our CRADA partners.

4. Discussion

4.1. Major findings

4.1.1. A generalized method for predicting serious drug AEs

The development of an *in silico* method for predicting serious hepatobiliary and urinary tract AEs of drugs represents a major advancement of this technology. (1) The experimental results reported in this study are based solely upon observations made in humans in pharmaceutical clinical trials and from data compiled in post-market surveillance databases maintained by FDA. The resulting QSAR models are not based on extrapolations from the results of animal studies. (2) Most serious AEs of pharmaceuticals are rare, idiosyncratic, and generally not detectable in clinical trials with relatively small numbers of patients (Navarro and Senior, 2006). That these effects can successfully be predicted using computational toxicology software could not necessarily be anticipated. However, this study demonstrated that roughly half of drugs showing serious hepatobiliary and urinary tract AEs that were missed in pre-market clinical trials could be correctly predicted using QSAR models. (3) The results demonstrate that QSAR technology provides a powerful tool that can be used to predict AEs of pharmaceuticals prior to human exposure in clinical trials and in post-market prescriptions. Although the mechanistic bases of hepatobiliary and urinary tract AEs are largely unknown, and few, if any, pharmaceutical chemical class warnings are available to the medical community, the *in silico* approach used in this study may also provide molecular insights into the mechanism(s) responsible for some AEs (Matthews et al., 2009).

Despite the potential utility of QSAR models to provide decision support information in drug discovery, lead chemical selection, and regulatory activities, little computational toxicological research has been done to identify global SARs for drugs having toxicologically related AE findings in humans. The only other investigation of this kind of which we are aware was an attempt by this laboratory to compile and construct QSAR models using a subset of SRS post-market AE data (Matthews et al., 2004). We believe that the use of QSAR programs to study human AEs will increase from this time forward.

4.1.2. QSAR models for predicting AEs are complementary

There is an increasing interest in consensus modeling and making use of batteries of QSARs (OECD, 2007; Matthews et al., 2008; Contrera et al., 2007; Votano et al., 2004). Our laboratory has also been interested in consensus modeling to provide decision support information for the FDA review process (Kruhlak et al., 2007). In the current study we compared the prediction paradigms of four QSAR programs employing a consensus prediction strategy using: (1) the same training data set in which drugs had the same AE activity scores; (2) the same LMO cross-validation method to assess program performance; and (3) the same QSAR model performance criterion (specificity ~80%).

The ability to control QSAR program specificity is considered to be particularly important because the specificity and sensitivity performance of QSAR models for each of the programs tested can be adjusted to meet the needs of the investigator. We chose a high specificity QSAR restriction to assess complementarity. The only experimental variable was the QSAR program prediction paradigm, which is markedly different for the four programs (Matthews et al., 2008). This strategy was employed because none of the QSAR programs was expected to exhibit both high specificity and high sensitivity, and it was anticipated that two or more of the programs might have complementary prediction paradigms.

In contrast, the historical approach to compare QSAR programs has been markedly different, e.g., the NTP exercises for predicting rodent carcinogenicity (Ashby, 1996; Ashby and Tennant, 1994); the EPA high production volume challenge program (US EPA, 1999); and other studies comparing global QSAR programs (Pearl et al., 2001; Tunkel et al., 2005). Unlike our study in which all of the experimental variables were controlled, these studies only employed the same set of test chemicals. These studies generally used: (1) different training data sets in which chemicals often did not have exactly the same toxicity scores, (2) no restrictions on QSAR model performance, and (3) testing in different laboratories. In such “bake off” competitions, the programs were compared and performance was rated by many investigators based only upon the predictive performance of the programs (e.g., Bager et al., 1997; Bristol et al., 1996). In all of these studies the differences in the performance of the programs was assumed to be due to the program's prediction paradigm alone and the investigation of consensus prediction strategies was limited.

Initially, our MC4PC QSAR models showed high specificity (90.5%), but they had low sensitivity (19.2%, data not presented). The performance of MC4PC and the three other computational toxicology programs was dramatically improved, however, by controlling two experimental parameters: (1) optimization of the method for classification of the drug AEs, and (2) identification of an optimal BP value for each AE endpoint. When these parameters were controlled, the four QSAR programs could, on average, predict drug AEs with a 39.3% sensitivity using the standard molecular fragment and molecular descriptor QSAR methodologies (Table 4). Furthermore, the sensitivity could be increased to 56.2% by combining any two of these programs (Table 6), or 67.9% by calling a chemical positive if predicted to be positive in at least one of four programs (Table 7).

All four programs were shown to be equally complementary (Table 5). None of the QSAR programs has a paradigm that completely models all the idiosyncratic relationships between AEs and pharmaceutical molecular structure; however, collectively the prediction paradigms for the four QSAR programs we tested offer a high confidence method for predicting serious drug AEs (Table 6). Even though we may not know the mechanisms by which drugs cause AEs, we can still predict some AEs based purely upon the molecular structure of the drug. The observation that combining complementary QSAR programs improves performance suggests that merging existing QSAR program algorithms could result in programs with substantially improved performance. MC4PC and LPDM molecular fragment prediction programs currently employ only a few molecular descriptors to make predictions; expansion of this functionality may be possible. Likewise, the MDL-QSAR and BioEpisteme molecular descriptor prediction paradigms now employ some connectivity indices, so expansion of this functionality may be feasible.

4.1.3. External validation study of pharmaceutical hepatotoxicity

The conduct of an external validation study is considered an important requirement for evaluation of the predictive performance of QSAR models (OECD, 2007). This validation method uses

test chemicals that were never used in the construction of the QSAR models being evaluated. It is considered by many investigators to be the most rigorous test of models (Perkins et al., 2003; Golbraikh and Tropsha, 2002). The only drugs not included in our AE database (Ursem et al., 2009) and QSAR models were those marketed outside the United States. We obtained a set of 18 drugs demonstrating serious hepatotoxicity that had been removed from the worldwide markets from 1960 to 1999 (Fung et al., 2001); these were not considered in our QSAR models.

The results of this external validation study (Table 8A) are comparable to the results of the consensus prediction experiment involving the entire database (Table 7). Under what we considered to be the best consensus strategy, requiring positive predictions by two or more QSAR programs for a chemical to be predicted positive, the sensitivity of the internal and external validation experiments were 46.3% (Table 7) and 74.1% (Table 8B), respectively. Alternatively, if all of the predictions from the four programs were accepted, the sensitivities of the internal and external validation experiments were 67.9% (Table 7) and 88.9% (Table 8B), respectively. Although the number of chemicals in the external validation experiment is small, these data clearly demonstrate that the QSAR methodology presented in this report is able to detect a substantial portion of the drugs that have exhibited severe hepatobiliary and urinary tract AEs.

4.2. Optimization of AE QSAR model experimental parameters

4.2.1. AE database development

The development of AE QSAR models has been hampered by several difficulties encountered in creating a unified AE database. Some of the problems encountered include non-uniform reporting of AEs across toxicological endpoints, patients often taking more than one medication at a time, and a higher AE reporting frequency during the initial period of drug marketing. It has also been a problem that the AERS and SRS database AE data are not linked to pharmaceutical chemical structures, and the SRS data are only linked to pharmaceutical trade names. Furthermore, merging the SRS and AERS data is not straightforward due to substantial differences in the way the data have been recorded. AEs from the AERS data sets are described using the Medical Dictionary for Regulatory Activities (MedDRA) terms (<http://www.meddrasso.com>), and approximately 14,000 terms are arranged hierarchically to reflect both organ systems and categories of toxic effects. However, the SRS database used the Coding Symbols for Thesaurus of Adverse Reaction (COSTART, 1995) vocabulary which has fewer than 1200 terms.

4.2.2. Low frequency of actives: Adjustment of A/I ratio

Development of QSAR models for pharmaceutical AEs has also been impeded because serious drug-related AEs are relatively rare, giving a low ratio of active to inactive drug molecules for most specific and composite AE endpoints. We have reported that a low percentage of actives, or a low A/I ratio, results in low sensitivity of MC4PC QSAR models for reproductive and developmental toxicity, and that the sensitivity is improved if the percentage of actives in the training data is adjusted by using the same active chemicals with different subsets of the inactive chemicals (Matthews et al., 2007b). Optimal MC4PC predictive performance was achieved when the percentage of actives in the training data set was 35–45%, which corresponds to an A/I ratio of 0.538–0.818. In the current study we were confronted by a comparable problem; all of the individual specific and composite hepatobiliary and urinary tract endpoints had <20% actives (Ursem et al., 2009). In studies done for this investigation, we found that the A/I adjustment procedure improved the sensitivity of the MC4PC, BioEpisteme and LPDM AE

QSAR models (Table 4). (In contrast, an adjustment of the A/I ratio had a detrimental effect on MDL-QSAR performance.)

4.2.3. Identification of optimal BP for active and inactive drug AEs

Pharmaceutical AEs form a continuum of PRR values over a 6-log range and there is no obvious BP that separates active and inactive drugs. Unfortunately, most QSARs designed for predicting toxicological responses require binary distributions of toxicology data in which active and inactive chemicals are clearly distinguished (e.g., mutagenic vs. non-mutagenic chemicals). Differences in the molecular properties of the active and inactive chemicals provide the basis for identification of toxicological activities related to specific chemical properties using a variety of *in silico* software programs. Because the boundary between drug-related human AEs and inactive drugs is poorly defined, it has been very difficult to identify clusters of structurally related drug molecules that cause AEs and to identify specific structural alerts for these drugs (which could become specific chemical class warnings for the medical community).

In an attempt to identify a BP for active and inactive drugs we constructed MC4PC AE QSAR models using a fixed BP value for all of the composite AE endpoints. The results of these experiments showed that acceptable QSAR models could be constructed for a few composite data sets, but that the majority of the QSAR models had unacceptably low sensitivity. In order to solve this problem and be able to make QSAR models with higher sensitivity for all of the composite endpoints, we identified individual BP values for each of the composite AE endpoints by determining empirically the optimal BP based upon the predictive performance of the QSAR program. Once again, this strategy is based upon our hypothesis that the best QSAR model performance occurs when the active and inactive drug molecules are correctly classified and optimally separated into two groups with the active molecules forming distinct clusters based upon shared chemical molecular properties, while poor predictive performance is hypothesized to be caused by misclassification of active and inactive molecules.

Since the four QSAR programs were shown to be complementary (Table 5), different optimal BP values for different QSAR programs were anticipated. The results of this study revealed that the optimal BP value for each of the 11 hepatobiliary and urinary tract endpoints for the MC4PC QSAR program (Column 4, Table 3) compared to the BioEpisteme program (Column 4, Table 2) was an average of 3.72 units higher (median = 4; range = 0–10) for MC4PC (Table 3). Furthermore, using an optimal BP value for MC4PC resulted in AE QSAR models that exhibited high predictive performance for all three of the remaining programs (Table 4). These data convinced us that a two-step method for identifying the BP value for each AE would be adequately sensitive for optimizing global QSAR prediction paradigms in which BioEpisteme would be used to make a preliminary estimate of the BP value, and then MC4PC would be used to determine a precise BP value that was adequate for the different QSAR program platforms.

4.2.4. PRR optimization studies

In this investigation the method used to identify drugs with significant AEs had a major effect on the predictive performance of QSAR models. We examined the impact of using different PRR criteria employed by pharmacovigilance groups to detect serious drug-related AEs to classify an AE response upon the predictive performance of the BioEpisteme QSAR program, including varying the number of reports required for determining a significant AE, using a χ^2 statistic vs. a fixed percentage of actives, and including a Yates correction of the χ^2 statistic. The results of these experiments showed that the highest sensitivities of the QSARs was obtained when a minimum requirement was set at 1–4 reports, with a 2 AE report minimum being optimal (Table 1). These data

also demonstrated that the use of the Yates correction for the χ^2 statistic consistently decreased the performance of the QSAR models. The observation that the Yates correction and high report minimum PRR criteria decreased the sensitivity of the QSAR is comparable to what we reported previously in our investigation of hepatotoxicity using only SRS AE data with shipping unit data used as a denominator for patient exposure. In that report, we used an AE report minimum of 2–3 reports to identify active drugs and construct the QSAR models (Matthews et al., 2004).

The results of the PRR optimization studies are very interesting to us because the experimental conditions that yielded the best QSAR model performance are much less stringent than conditions now employed by pharmacovigilance groups and FDA/CDER for regulatory decisions. Pharmacovigilance groups require that a significant drug AE has 3 or more AE reports for different patients, a χ^2 of 4, and a Yates correction of the χ^2 value. FDA/CDER is even more stringent, using an empirical Bayesian, multi-term gamma poisson shrinker analysis (DuMouchel, 1999; Almenoff et al., 2006; Szarfman et al., 2003, 2004). The most likely explanation for the difference in these findings is that the PRR values suitable for QSAR model construction have a requirement that the active drugs share structural properties that are recognized by the QSAR program prediction paradigm and form representative clusters based upon these properties. If the QSAR programs cannot identify clusters of actives with significant molecular properties, the QSAR models will exhibit poor predictive performance. In contrast, the highest priority for the pharmacovigilance approach is to make a very high confidence prediction of what drugs are active and gives a lesser priority to identification of molecular features of the active drug.

4.2.5. Identification of the best source of post-market surveillance AE data

At the beginning of this investigation we developed separate MC4PC QSAR models for the specific AE endpoints in the SRS and AERS databases. This decision was made because the two AE databases use different AE vocabularies and different rules for reporting the number of medications and AEs in a patient report. The concern was that AE data derived from the different sources might have an effect on the performance of the models. However, early studies revealed that the source of the AE data had only a small effect on the sensitivity of the models and that, in general, the larger the training data set for the AE model, the better the predictive performance of the model. This latter result provided evidence that our procedure for merging the AERS and SRS AE vocabulary terms and pooling pharmaceutical trade names was sound and did not cause significant artifacts in the QSAR models.

In these same studies we also evaluated the best method for merging the AERS and SRS data sets for drugs with AEs in both databases. The sensitivity of models containing the larger of the AERS or SRS activity unit values was found to be consistently higher than models using the average of the SRS and AERS value. This result could be related to the observation that the majority of pharmaceutical AEs are reported during the first five years of marketing (Szarfman et al., 2004). The majority of the highest activities occurred in the SRS database and the SRS contained the first year AE reports for the majority of the drugs in the two databases (except for those that were marketed in the mid- to late-1990s).

4.2.6. Use of published AE data

Throughout this investigation we have sought the best way to utilize AE data from all possible sources, including AEs reported in clinical trials, post-market surveillance, and the literature. Although AEs reported in human clinical trials are carefully conducted experiments with a denominator for patient exposure, the small number of patients in the trials limits their sensitivity for

detection of some serious drug AEs. Thus, the results of clinical trials that were reported in the literature were considered in this investigation, but the data from these trials provided very few serious AEs. In the end we found that AEs reported in post-market surveillance and in the literature to be the most useful in QSAR model construction.

Because the literature is a robust source of AE data for pharmaceuticals that is based upon both clinical studies and anecdotal observations in patients, we attempted to find the best way to incorporate these AE data into our QSAR models. Although there was substantial agreement in the activities of drugs based upon analyses of the AERS and SRS data sets and literature reports, there was a small number of drugs with significant AE findings reported in the literature that were inactive based upon analyses of the AERS and SRS data (Ursem et al., 2009). In order to incorporate the drugs with significant AE findings into our models, we selected only those drugs that had been removed from the market due to these specific AEs, or when the AEs were observed in >2% of the patients receiving the medication. We speculate that there may be two reasons for literature AE actives. There may have been under-reporting of certain AEs when drug-related findings are reported in the literature and commonly known to physicians. Alternatively, drugs that were approved in Europe and later removed from the market, were often not marketed in the US and pharmacovigilance may be under-reported in the AERS and SRS databases.

When actives based upon the published literature were added to the QSAR models, there usually was a substantial improvement in the performance of the QSAR models. Four composite endpoints showing substantial improvement when literature data of this sort were considered were the jaundice/cholestasis, liver enzymes, kidney function tests models, and bladder disorders. Nevertheless, in some experiments the addition of literature data had no obvious impact on QSAR model performance (Table 1). It should be emphasized that we only added data from the published literature on active drugs when there was substantial evidence of drug-related AE findings. Because we did not observe any AE endpoint QSAR models that were compromised by the addition of literature AE data, we recommend limited consideration of these data for QSAR model supplementation.

4.3. Possible limitations in the investigation

■ The current QSAR models are based on about 1600 drugs with hepatobiliary and urinary tract disorder AEs. It is likely that larger training data sets would exhibit improved predictive performance. Although the current investigation included all of the pharmaceuticals marketed in the United States from 1969 through 2006, there are two small groups of drugs missing from the current QSAR models: (1) drugs that were approved and marketed outside of the US and having substantial AE clinical trial and post-market surveillance data and (2) new molecular entities approved and marketed after June, 2006. Thus, the current QSAR models might be enhanced with small numbers of drugs, but a large increase in the data sets is not likely.

■ We may have underestimated the number of drugs having significant AE findings based upon the published literature when these drugs were previously scored as inactive based upon AERS and SRS data.

■ The current investigation included 7,342,676 AE reports (~20,000,000 drug/AE records) from AERS and SRS databases entered through June, 2006. Since about 1000 AE reports are added each day, enhancement of the QSAR model training data sets will be useful at some point in the future. It is possible that the addition of additional AE data might change the PRR activity values of some drugs in our current QSAR models. These changes would more

likely occur for newer drugs for which, in general, more reports are received.

■ There were certain specific AE data sets with a small number of reports that were not related to the 11 composite AE endpoints included in this study. If a specific AE data set has fewer than 5% actives, it is difficult to construct a QSAR model, even when the A/I ratio is adjusted. The *in silico* methodology described in this report is currently inappropriate for very rare AE endpoints.

■ In this investigation we determined an optimal BP to separate active and inactive drugs based upon the predictive performance of QSAR models. It is possible that better methods to define the BP will be suggested.

■ The *in silico* methodology described in this report is limited to simple organic chemicals and is not suitable for chemical classes not covered by the applicability domain of the QSAR models.

■ The coverage of the QSAR models was shown to be very high for drug-like molecules (Table 4) and is anticipated to be good for other FDA-regulated products (e.g., food additives, natural plant substances, etc.). However, the QSAR models may have difficulty predicting industrial chemicals.

■ Many of the QSAR modules in this study exhibited low sensitivity (Table 4), and this may be problematic for some applications of this methodology. The low sensitivity issue can be partially overcome through the use of consensus prediction strategies. Because all four QSAR programs were shown to be complementary (Table 5), inclusion of all predictions from two or more QSAR programs substantially improves the sensitivity of the QSAR models (Tables 6 and 7). Nevertheless, there seems to be a significant portion of drug AEs that are not predicted by any of the four programs. We believe that future innovations in the technology will further improve the sensitivity of the models.

■ MDL-QSAR is a toolbox of different QSAR methodologies. Using others not employed for this study might improve the performance of AE QSAR models.

■ In this investigation we relied only upon the BioEpisteme default prediction paradigm; thus, substantial improvements in AE models may be possible. For example, no attempt was made to optimize the predictive performance of the individual AE QSAR models by eliminating additional descriptors that might diminish the performance of the models. It is anticipated that this approach could improve the performance of the models. The current study used the default genetic algorithm to predict AEs, but the BioEpisteme platform has several algorithms available that could be tested. Furthermore, previous experiments in this laboratory with a rodent carcinogenicity mechanism of drug action QSAR models based on 3D descriptors suggest the 3D descriptor functionality should be helpful.

■ In this investigation we relied only upon the default LPDM prediction paradigm; substantial improvements in AE models may be possible. For example, we made no attempt to optimize the predictive performance of the individual AE QSAR models by manually substituting different rules and fingerprint scaffolds that might improve performance. It is also possible to expand the expert rule functionality to include substantial libraries of 2D, 3D, and 4D molecular descriptors, as well as additional fingerprint libraries.

■ Other state-of-the-art QSAR programs which are currently available might demonstrate higher predictive performance than the four programs we tested. However, due to the limited resources available to this applied research group, investigations with additional global QSAR prediction programs have not been feasible.

■ In Section 3.10 we hypothesize that the higher confidence, consensus positive predictions obtained using 2, 3, or 4 QSAR programs might preferentially identify drugs that had been assigned a higher WOE for AEs. However, the results of this study showed that

the subset of active drugs having consensus AE predictions did not have higher activities compared to active drugs that were not predicted positive (column 3, Table 7). This means that the biological factors that were used to create high activity scores for AEs are not always correlated on a one-to-one basis with the QSAR prediction paradigms employed in this study.

Possible explanations for this phenomenon could be that: (1) the majority of pharmaceuticals scored as actives have only borderline (marginal) AE activity; (2) the assignments of actives vs. inactives are possibly flawed; (3) the methodologies for QSAR model building were possibly flawed, and/or (4) the biological complexity of idiosyncratic drug-related AEs may not be adequately addressed by the QSAR prediction paradigms employed in this study. The first two possibilities cannot be excluded, but are regarded as unlikely because the active and inactive classifications were statistically very stringent, based upon a PRR value of 2.0 and χ^2 of >4 ($p < 0.05$), and the entire AE database is very large (>20,000,000 drug/AE records). The third possibility also cannot be excluded, but it is considered unlikely because the results of the internal cross-validation experiments, and the single external validation experiment, demonstrate that the QSAR models for the AEs are working as well as could be expected for a training data set of only 1600 chemicals. Thus, we believe discordance between experimental activity and QSAR predictions is related to the QSAR paradigms not fully representing the full biological complexity of idiosyncratic drug-related AEs. For example, there may be hundreds of MOAs by which drugs cause AEs and our QSAR paradigms only represent a portion of these MOAs. The answer to this issue may come through resolving one or more of the possible limitations in the study discussed above.

■ In Section 3.11 we presented the results of a small external validation study, but it did not contain inactive drugs which are normally included in an external experiment. At the time this study was conducted all of the inactive drugs marketed in the US were included in the QSAR models. To overcome this concern, it is the intent of this laboratory to conduct a prospective study of the performance of the hepatobiliary, urinary tract, and other organ system QSAR models developed in the future. We intend to use drug warning and labeling information that was recently reported in FDA/CDER's MedWatch Program (June 2006 to present), which were not considered in the current investigation. The results of this prospective study will be reported in a subsequent publication from this laboratory.

The results of the external validation study were also limited in that they did not attempt to connect the individual predictions from the AE models with the specific type of AE that resulted in the drugs being removed from the market. The purpose of the QSAR models developed in this investigation is to provide a WOE that a given pharmaceutical may have a profile of significant AEs for an organ system. Furthermore, the withdrawn drugs often had more than one significant hepatobiliary AE besides liver failure that was documented in a small number of patients and was the primary reason for their removal from the market.

4.4. Future investigations

■ None of the QSAR programs in the current investigation utilize 3D descriptors to predict AEs. Future studies may reveal that AEs are related to the pharmacological properties of drugs and thereby would be better detected using 3D and 4D prediction strategies.

■ The current investigation did not explore the possibility that a substantial portion of the serious drug AEs are related to drug metabolism. Since the major metabolites of most drugs in humans are known, it is entirely feasible to investigate relationships between drug metabolites and AEs.

■ In the near future, MC4PC may be able to examine the relative reliability and consistency of the individual chemical toxicological activities in a QSAR model. It is possible that this functionality could improve QSAR model performance by identifying drugs that might be more appropriately classified as marginals or removed from the model entirely.

■ We will continue to explore the effects of reclassification of drugs with marginal AE findings as actives on QSAR model predictive performance. The best QSAR model performance to date has been achieved by classification of marginals as inactive. However, it is possible that an additional experimental parameter could be used to partition marginal drug activities into actives and inactives and identify a subset that are truly active.

■ This investigation focused on the development of hepatobiliary and urinary tract AE QSAR models. Investigations are already in progress to expand the AE endpoints to include additional organ systems including heart, endocrine, respiratory, and immune system disorders.

■ This investigation tested several different PRR methods of scoring AEs, but some methods were not tested. Future investigations are planned to evaluate the MPCS system currently employed by FDA (Almenoff et al., 2006; DuMouchel, 1999). This method stratifies the AEs for numerous factors and may be useful in elimination of some false positive drugs detected by the current AE QSAR models.

■ This investigation was focused on the prediction of serious drug AEs. Parallel investigations on relating these AEs to specific mechanisms of action of drugs is addressed in part C of this investigation (Matthews et al., 2009).

■ This investigation used almost all of the available AE data then available to us to construct models with the greatest possible applicability domain to maximize the predictive performance in standard LMO validation experiments. We intend to conduct a standard external validation experiment in the future with a larger and more representative test set of compounds as new data become available.

4.5. Applications for the AE QSARs

The optimized hepatobiliary and urinary tract disorder QSAR models described in this report are being used internally within FDA to provide decision support information for a variety of regulatory and research applications. These same models will be made available through our CRADA partners, subject to appropriate approval, to allow users in the scientific community to obtain the same prediction results as are obtained by FDA. At the FDA, the QSAR models are used to provide information on contaminants in drug preparations and known metabolites of pharmaceutical active ingredients. They are also being used in prospective studies as a proactive method to provide useful decision support information for the early detection of pharmaceutical AEs in humans. We expect the QSAR models may provide a rapid and effective means of prioritizing AE concerns for drugs, and lead to identification of pharmaceutical chemical class warnings. We anticipate that these AE models will also be used to provide decision support information for lead chemical selection activities under agreements between FDA and the National Institutes of Health and regulatory activities of the Environmental Protection Agency (EPA).

This article may provoke serious discussion of the use of global QSARs for adverse human effects to provide decision support information for regulatory decisions and lead selection and discovery applications in industry. Although the use of multiple QSAR programs offers clear predictive advantages over the use of a single program, this approach creates some uncertainties for the scientific community. For example, what is the ideal combination of *in*

in silico tools or QSAR and expert system algorithms for consensus modeling? Should the user of tools developed at the FDA accept the default QSAR model predictions or create alternative models? Most users will want to know the optimal way to evaluate the *in silico* data to meet their own needs and applications.

In anticipation of this discussion, the FDA is engaged in a three phase process to implement this *in silico* technology. The first phase is the development of specific global QSAR and expert system methodologies that offer rapid and reliable decision support information to meet FDA regulatory and research needs and to make these tools available to the scientific community. Scientists need tools that provide an *in silico* dossier of information on both chemical toxicities for endpoints that cannot be addressed in humans (carcinogenicity, genetic toxicity, and reproductive and developmental toxicity; Matthews et al., 2006a,b, 2007a,b, 2008), as well as adverse effects of chemicals in humans that are detected in clinical trials and post-market surveillance databases (Matthews et al., 2009). Updates to the tools will become available as data are harvested from Agency archives, data sharing efforts with other agencies are implemented, and the training data sets of the *in silico* tools are enhanced. The second phase is to introduce, educate, and build consensus support for this new global QSAR and expert system testing paradigm, and the FDA is currently involved in this process. The third phase, soon to commence, involves seeking public comment on FDA Guidances for the proper conduct of these *in silico* studies in support of regulatory applications. Once in place, these *in silico* methods will be performed as a means of reduction, replacement, refinement for longer, more expensive testing, not as an additional burden.

The FDA approach is modeled after the US EPA which has been a pioneer in the development of QSAR technology to provide decision support information for EPA regulatory decisions and applications. The EPA has developed QSAR and expert system tools to predict environmental fate (EPI Suite and PBT Profiler), chemical toxicity (ECOSAR and Oncologic Cancer Expert System), exposure (Chem-STEER and E-fast), and chemical structure analogs with data in publicly available databases (AIM) (<http://www.epa.gov/oppt/sf/tools/methods.htm#new>). Because these tools are often based upon knowledge derived from proprietary studies, and the confidential business information can not be disclosed, these tools were prepared by the EPA. Similarly, the EPA wanted to ensure that anyone using these tools would obtain the same prediction; thus, the software tools were developed as read only applications. The EPA uses established policy and procedures for the development and verification of the tools and models, and it makes available to the public information guidelines, guidances, software training, as well as the software.

Acknowledgments

The research described in this investigation would not have been possible without the extensive contributions by our collaborators.

References

- Almenoff, J.S., LaCroix, K.K., Yuen, N.A., Fram, D., DuMouchel, W., 2006. Comparative performance of two quantitative safety signaling methods. *Drug Saf.* 29, 875–887.
- Ashby, J., 1996. Prediction of rodent carcinogenicity for 30 chemicals. *Environ. Health Perspect.* 104, 1101–1104.
- Ashby, J., Tennant, R.W., 1994. Prediction of rodent carcinogenicity for 44 chemicals. *Mutagenesis* 9, 7–15.
- Bager, Y., Lindebro, M.C., Martel, P., Chaumontet, C., Warngard, L., Benigni, R., 1997. The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. 4. Prediction of non-lethal mammalian toxicological and points, and expert systems for toxicity prediction. *Mutat. Res.* 387, 35–45.
- Benigni, R., 2005. Structure activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem. Rev.* 105, 1767–1800.
- Benz, R.D., 2007. Toxicological and clinical computational analysis and the US FDA/CDER. *Expert Opin. Drug Metab. Toxicol.* 3, 109–124.
- Bristol, D.W., Wachsmann, J.T., Greenwell, A., 1996. The NIEHS predictive-toxicology evaluation project. *Environ. Health Perspect.* 104, 1001–1010.
- Contrera, J.F., Kruhlak, N.L., Matthews, E.J., Benz, R.D., 2007. Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR Models. *Regul. Toxicol. Pharmacol.* 49, 172–182.
- Cooper, J.A., Saracci, R., Cole, P., 1979. *Br. J. Cancer* 39, 87–89.
- COSTART. 1995. Coding Symbols for Thesaurus of Adverse Reaction Terms, fifth ed., Department of Health and Human Services: Rockville, MD.
- DuMouchel, W., 1999. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. *Am. Stat.* 53, 177–202.
- Evans, S.J., Waller, P.C., Davis, S., 2001. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* 10, 483–486.
- Fielden, M.R., Eynon, B.P., Natsoulis, G., Jarnagin, K., Banas, D., Kolaja, K.L., 2005. A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity. *Toxicol. Pathol.* 33, 675–683.
- Fung, M., Thornton, A., Mybeck, K., Wu, J.H., Hornbuckle, K., Muniz, E., 2001. Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets-1960 to 1999. *Drug Inform. J.* 25, 293–317.
- Golbraikh, A., Tropsha, A., 2002. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des.* 16, 357–369.
- Gombar, V.K., Mattioni, B.E., Zwickl, C., Deahl, T., 2007. Computational Approaches for Assessment of Toxicity—A Historical Perspective and Current Status. In: Elkins, S. (Ed.), *Computational Toxicology: Risk Assessment for Pharmaceuticals and Environmental Chemicals*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 183–215.
- Hanley, J.A., 1989. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit. Rev. Diagn. Imaging* 29, 307–335.
- Kruhlak, N.L., Contrera, J.F., Benz, R.D., Matthews, E.J., 2007. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv. Drug Deliv. Rev.* 59, 43–55.
- Lee, W.M., 2003. Acute liver failure in the United States. *Semin. Liver Dis.* 23, 217–226.
- Lee, W.M., 2005. Drug-induced acute liver failure in the US 2005: results from the US acute liver failure study group. Available from: <http://www.fda.gov/cder/livertox/presentations2005/William_Lee.ppt>.
- Lusted, L.B., 1971. Signal detectability and medical decision-making. *Science* 171, 1217–1219.
- Matthews, E.J., Contrera, J.F., 1998. A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. *Regul. Toxicol. Pharmacol.* 28, 242–264.
- Matthews, E.J., Contrera, J.F., 2007. In silico approaches to explore toxicity endpoints: Issues and concerns for estimating human health effects. *Expert Opin. Drug Metabol. Toxicol.* 3, 125–134.
- Matthews, E.J., Kruhlak, N.L., Benz, R.D., Contrera, J.F., Marchant, C.A., 2008. Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows software to achieve high performance, high confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol. Mechan. Methods* 18, 189–206.
- Matthews, E.J., Kruhlak, N.L., Benz, R.D., Contrera, J.F., Marchant, C.A., David Aragonés Sabaté. 2009. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: C: Use of (Q)SAR and Expert systems for estimation of the mechanism of action of drug-induced urinary tract and hepatobiliary toxicities. *Regul. Toxicol. Pharmacol.*, in press, doi:10.1016/j.yrtph.2009.01.007.
- Matthews, E.J., Kruhlak, N.L., Benz, R.D., Contrera, J.F., 2007a. A comprehensive model for reproductive and developmental toxicity hazard identification: I. Development of a weight of evidence QSAR database. *Regul. Toxicol. Pharmacol.* 47, 115–135.
- Matthews, E.J., Kruhlak, N.L., Benz, R.D., Ivanov, J., Klopman, G., Contrera, J.F., 2007b. A comprehensive model for reproductive and developmental toxicity hazard identification: II. Construction of QSAR models to predict activities of untested chemicals. *Regul. Toxicol. Pharmacol.* 47, 136–155.
- Matthews, E.J., Kruhlak, N.L., Cimino, M.C., Benz, R.D., Contrera, J.C., 2006a. An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints. *Regul. Toxicol. Pharmacol.* 44, 83–96.
- Matthews, E.J., Kruhlak, N.L., Cimino, M.C., Benz, R.D., Contrera, J.F., 2006b. An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. Identification of genotoxicants, reprotoxicants, and carcinogens using in silico methods. *Regul. Toxicol. Pharmacol.* 44, 97–110.
- Matthews, E.J., Kruhlak, N.L., Weaver, J.L., Benz, R.D., Contrera, J.F., 2004. Assessment of the health effects of chemicals in humans: II. Construction of an adverse effects database for QSAR modeling. *Curr. Drug Discov. Technol.* 1, 243–254.
- Moore, N., Thiessard, F., Begaud, B., 2005. The history of disproportionality measures (reporting odds ratio, proportional reporting rates) in spontaneous reporting of adverse drug reactions. *Pharmacoepidemiol. Drug Saf.* 14, 285–286.
- Navarro, V.J., Senior, J.R., 2006. Drug related hepatotoxicity. *N. Engl. J. Med.* 354, 731–739.

- OECD, 2007. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD environment health and safety publications series on testing and assessment No. XX. Draft.
- Pearl, G.M., Livingston-Carr, S., Durham, S.K., 2001. Integration of computational analysis as a sentinel tool in toxicological assessments. *Curr. Top. Med. Chem.* 1, 247–255.
- Perkins, R., Fang, H., Tong, W., Welsh, W.J., 2003. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* 22, 1666–1679.
- Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. *Machine Learn. J.* 42, 203–231.
- Szarfman, A., Machado, S.G., O'Neill, R.T., 2003. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf.* 25, 381–392.
- Szarfman, A., Tønning, J.M., Doraiswamy, P.M., 2004. Pharmacovigilance in the 21st century: new systematic tools for an old problem. *Pharmacotherapy* 24, 1099–1104.
- Temple, R.J., 2001. Hepatotoxicity through the years: impact on the FDA. Available from: <<http://www.fda.gov/cder/livertox/presentations/im1389/sld001.htm>>.
- Tunkel, J., Mayo, K., Austin, C., Hickerson, A., Howard, P., 2005. Practical considerations on the use of predictive models for regulatory purposes. *Environ. Sci. Technol.* 39, 2188–2199.
- Ursem, C.J., Matthews, E.J., Kruhlik, N.L., Contrera, J.F., Benz, R.D., 2009. Identification of structure activity relationships for adverse effects of pharmaceuticals in humans A: Use of FDA post market reports to create a database of hepatobiliary and urinary tract toxicities. *Regul. Toxicol. Pharmacol.*, in press, doi:10.1016/j.yrtph.2008.12.009
- US EPA, Office of Pollution Prevention and Toxics, 1999. The use of structure-activity relationships (SAR) in the high production volume chemicals challenge program. Available from: <<http://www.epa.gov/chemrtk/sarfin1.htm>>.
- Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., 2004. 3 new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* 19, 365–377.
- Zimmerman, H.J., 1978. Drug-induced liver disease. *Drugs* 16, 25–45.
- Zimmerman, H.J., 1999. Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver, second ed. Lipincott Williams and Wilkins, Philadelphia.