Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jona Marie Ogden Date

Using Quantitative Structure-Activity Relationships (QSAR) to Establish Toxicity/Environmental Scores (TES)

By

Jona Marie Ogden
MPH

Department of Environmental Health

_____
W. Michael Caudle, PhD
Committee Chair

_____
Patricia Ruiz, PhD
Committee Member

_____
Paige Tolbert, PhD
Committee Member

Using Quantitative Structure-Activity Relationships (QSAR) to Establish Toxicity/Environmental Scores (TES)

By

Jona Marie Ogden
B.S., Environmental Health Science
The University of Georgia
2009

Thesis Committee Chair: W. Michael Caudle, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Environmental Health
2012

Abstract

Using Quantitative Structure-Activity Relationships (QSAR) to Establish Toxicity/Environmental Scores (TES)
By Jona Marie Ogden

The Agency for Toxic Substances and Disease Registry (ATSDR) uses Reportable Quantities (RQs) established by the Environmental Protection Agency (EPA) in order to prioritize substances subject to Toxicological Profile development. RQs are calculated using two distinct criteria. The first criteria is based on the intrinsic physicochemical (ignitability/reactivity) and toxicological properties (aquatic toxicity, acute mammalian toxicity, chronic toxicity, and potential carcinogenicity) of each chemical. The second criteria is based on a chemical's susceptibility to biodegradation, hydrolysis, and photolysis (BHP). When an RQ is not available, ATSDR uses the same criteria to develop a Toxicity/Environmental Score (TES). Sufficient original data are not available to assign a TES to many candidate chemicals. However, Quantitative Structure-Activity Relationship (QSAR) methods can be used to computationally predict the physicochemical, toxicological and biodegradability properties needed to calculate TESs. To evaluate the potential use of QSAR methods to estimate TESs, the physicochemical, toxicological and biodegradability properties of 102 chemicals were computationally-predicted, and QSAR TESs estimated. QSAR rat oral LD_{50}, fathead minnow LC_{50}, and BHP models predicted TESs that correlated strongly (71%, 53%, and 67%, respectively) with original TESs. QSAR could not predict a dose-response relationship needed to score chronic toxicity. However, an alternate approach combining developmental toxicity and chronic LOAELs was used to estimate chronic toxicity values. Using 1 of 4 proposed methods, QSAR-derived TESs were identical to original TESs for 57% of the chemicals evaluated. 89% of predicted TESs were within 1 tier of original TESs. Thus, QSAR methods may be used as an alternative approach to fill in data gaps needed for calculation of TESs. To optimize the use of *In Silico* prediction, an integrated approach for the use of multiple QSAR models, tools and approaches is needed.

Using Quantitative Structure-Activity Relationships (QSAR) to Establish Toxicity/Environmental Scores (TES)

By

Jona Marie Ogden
B.S., Environmental Health Science
The University of Georgia
2009

Thesis Committee Chair: W. Michael Caudle, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Environmental Health
2012

Table of Contents

List of Tables

List of Figures

List of Formulas

INTRODUCTION

In response to the congressional mandate in The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) of 1980, as amended by the Superfund Amendments and Reauthorization Act (SARA) of 1986, the Environmental Protection Agency (EPA) and Agency for Toxic Substances and Disease Registry (ATSDR) prepare and revise a priority list of hazardous substances most commonly found at facilities on the CERCLA National Priority List (NPL) every two years and develop Toxicological Profiles. Toxicological Profiles are comprehensive documents that detail a substance's toxicological properties (ATSDR, 2008). In keeping with the mandate, ATSDR uses the Priority List of hazardous substances, named the Substance Priority List (SPL), to prioritize toxicological profile development and, subsequently, a candidate for the identification of priority data needs (ATSDR, 2011a). The SPL includes hazardous substances that have been determined to be of greatest public health concern to persons at or near NPL sites. The first SPL, published in 1987, was comprised of 100 substances and expanded to include 847 hazardous substances by 2011.

Ranking methodology is used to generate the SPL. Ranking of substances on the SPL is based on three component scores that are summed to establish the total score. Total scores are calculated using Formula 1:

Formula 1. Formula for Substance Priority List Ranking

$$TOTAL\,SCORE = NPL\,FREQUENCY + TOXICITY + POTENTIAL\,FOR\,HUMAN\,EXPOSURE$$

(1,800 *max. points*) (600 *points*) (600 *points*) (300 *conc. pts.)* + (300 *exposure pts.)*

The three components of the total score are frequency at NPL sites, toxicity, and the potential for human exposure to the substance. The toxicity of a substance accounts for one third of the substance's total score and, thus, is important to the substance's overall ranking. EPA and ATSDR use the Reportable Quantity (RQ) approach as the quantitative measure of toxicity in the

total score algorithm. RQs are regulatory numbers established by EPA. If the RQ is exceeded, hazardous substance releases must be reported to local, state, and national authorities (ATSDR, 2011a). The toxicity data used in the RQ approach is derived from primary peer-reviewed literature. RQs have already been established for the majority of hazardous substances that are frequently detected at hazardous waste sites. Moreover, the determination of RQ health effect values uses weight-of-evidence considerations in evaluating data.

The comprehensive approach used to establish RQS makes them strong indicators of substance toxicity. RQs are established in a two-step process (EPA, 2011). The first step is an evaluation of intrinsic physical, chemical, and toxicological properties, also known as primary criteria. These include acute toxicity, chronic toxicity, carcinogenicity, and aquatic toxicity. Substance-specific rat oral lethal dose (50% kill) (LD_{50}) data, when available, is used to estimate acute toxicity. Substance-specific lethal concentration (50% kill) (LC_{50}) data for fathead minnows (Pimephales promelas) or bluegills (Lepomis macrochirus), when available, is used to estimate aquatic toxicity. Substance-specific mammalian Minimum Effective Doses (MED) are adjusted by species to derive human equivalent doses and multiplied by a rating value (RV_d) based on type of effect to estimate human chronic toxicity. Carcinogenicity data are based on rat or mice studies performed by the EPA or the International Agency for Research on Cancer (IARC). Substances evaluated for carcinogenicity are scored “high,” “medium,” or “low” based on either EPA or IARC cancer classification and given a primary criteria RQ of 1, 10, or 100, respectively. Each criteria value is assigned to one of five tiered RQ categories (1, 10, 100, 1000, or 5000), and the lowest RQ among all criteria is selected as the primary criteria RQ for the substance.

The second step is to evaluate substances for hydrolysis, photolysis, and biodegradation. If by one of these processes a substance degrades rapidly in the environment to a less harmful form, then the substance’s primary criteria RQ is raised one tier, establishing a final RQ. If the substance degrades to a more

harmful form, then the substance is assigned the final RQ of the more hazardous substance.

When an RQ is not available, ATSDR uses the same criteria and five tiered categories to develop an equivalent Toxicity/Environmental Score (TES) as a surrogate for the toxicity component in the total score algorithm. These scores are developed for use only in the ranking methodology and do not represent regulatory values. TESs have been assigned to more than 450 candidate hazardous substances. Hazardous substances that received a TES greater than 5,000 using the RQ methodology were dropped to the bottom of the candidate list because of their lack of known toxicity, and they received a total score of zero points.

However, data is not available to establish RQs and TESs for all substances found at NPL sites. Currently, approximately 3,400 uniquely identifiable hazardous substances have been identified at hazardous waste sites according to the ATSDR database. However, many candidate substances have data gaps that must be filled in order to assign an RQ or TES. Even substances assigned an RQ/TES may lack experimental data for one or more criteria.

In silico models have been used by academia, pharmaceutical, agrochemical, food, and other industries, as well as by various advisory and regulatory government agencies as decision support tools to fill data gaps in the toxicity data base of a substance (Cronin et al., 2003; el-Masri, Mumtaz, Choudhary, Cibulas, & De Rosa, 2002). These tools allow for the assessment of substances for which no data are available. Structure Activity Relationships (SARs) and Quantitative Structure Activity Relationships (QSARs) are increasingly being used as core prediction systems in toxicology (Barratt, 1998; McKinney, Richard, Waller, Newman, & Gerberick, 2000). Studies of SAR/QSAR have proven to be powerful tools to increase our understanding of the potential harmful effects of substances on the environment and human health.

In silico toxicology is an applied science that integrates mathematics, biology, chemistry, and computer technology to enable researchers to assess a substance's potential toxicological activity when experimental toxicological data

are lacking (el-Masri, et al., 2002; Mumtaz et al., 1995; Ruiz, Mumtaz, & Gombar, 2011). SAR/QSAR approaches mathematically correlate a substance structure's molecular attributes to its physicochemical, biological or toxicological activity (Accelrys, 2006; Ruiz, et al., 2011; Rupp, Appel, & Gundert-Remy, 2010). As the number and variety of potentially hazardous substances continues to increase, regulatory authorities have approached the challenge by applying *In silico* tools such as SAR/QSAR as guidance and/or decision support to substance risk-based approaches (Demchuk, Ruiz, Chou, & Fowler, 2011). Furthermore, the ever-increasing economic, social, and political call to reduce animal testing in toxicity evaluation has led to an expansion of the use of these tools. It is extensively documented that there are many different *In silico* SAR/QSAR models and platforms for predicting a wide range of toxicological endpoints. Many of these are commercially available (e.g. MultiCASE, DEREK, TOPKAT®), others are open sources (e.g. OncoLogic™, ToxCast™, ECOSAR, OASIS) and some are proprietary in-house systems (e.g. FDA QSAR models) (Demchuk, et al., 2011).

For the present study, the commercially available QSAR software, Toxicity Prediction by Komputer Assisted Technology (TOPKAT) was used. This software generates toxicity predictions based on a substance's structural similarity to large data sets of toxicological information retrieved from literature and stored in TOPKAT's database (Accelrys, 2006). The QSAR tool was able to assess various endpoints, including those used to assign an RQ/TES (weight of evidence/carcinogenicity, rat oral LD_{50}, chronic lowest observable adverse effect level (LOAEL), probability of biodegradability, and fathead minnow LC_{50}), when experimental data were limited or unavailable. The QSAR model can be used to assess the five criteria needed to develop provisional QSAR TESs for substances that lack the criteria.

The purpose of this study is to assess the application of SAR/QSAR models to develop surrogate/provisional QSAR TES values using an *In silico* approach. The QSAR model discussed in this study will produce surrogate TES values that will allow scientists to rank substances for toxicological profile

development when experimental data are not available. It will also allow risk assessors to make decisions based on a substance's predicted toxicity though experimental data are insufficiently available.

METHODS

DATASET

There were 847 substances in the 2011 SPL. All substances assigned a TES were selected from the 2011 SPL. Because SAR/QSAR approaches cannot be used to assess metals, salts, radio nuclides, polymers, and mixtures, these substances were excluded from the dataset. QSAR approaches were able to assess 293 of the 847 substances on the SPL. However, 191 of these substances were excluded because they fell outside TOPKAT's Optimum Prediction Space (OPS). The remaining 102 substances met the criteria for and were evaluated in this study.

Simplified Molecular Input Line Entry Specification (SMILES) codes for the remaining 102 substances were entered into the software to assess the following endpoints: carcinogenicity (Weight of Evidence), acute toxicity (rat oral LD_{50}), rat oral chronic toxicity (chronic oral LOAEL), aquatic toxicity (fathead minnow LC_{50}), and biodegradability (BHP).

QSAR PROTOCOL

TOPKAT offers three separate tests to evaluate the reliability of a prediction. The first test checks for substance substructures that were not evaluated during TOPKAT model development (Accelrys, 2006). Toxicity predictions for substances with substructures not accounted for in the existing TOPKAT database model are not considered to be reliable. The second test is an evaluation of substance descriptor values to determine if they are in the range of the descriptor values in the QSAR model database. The third test is an evaluation of substances to determine if they lie within the model's OPS. Predictions outside of a model's OPS are not supported by the model and thus deviate considerably from experimental values. Any substance that failed to meet

the criteria for all three tests for all models was removed from the dataset because of expected unreliability (n=191).

Carcinogenicity

The QSAR carcinogenicity model predictions were based on an overall weight of evidence (WOE) using a model that combined datasets from the National Toxicology Program (NTP) and the Food and Drug Administration (FDA) (Accelrys, 2006). This model was comprised of one statistically-significant module. The cross-validated accuracy of the model is shown in Table 1.

Table 1. Cross-validated accuracy of the carcinogenicity model[a]

Chemical Class	Number of Compounds	Specificity (%)	Sensitivity (%)	Indeterminate (%)
Aliphatics	112	93	93	0
Single & Multiple Benzenes	244	95	95	7
Heteroaromatics	127	97	95	3

[a](Accelrys, 2006)

EPA assigned an RQ for substance carcinogenicity using an approach that required a dose-response relationship. Such relationships were not available computationally because the QSAR model only predicted probabilities for carcinogenicity. Therefore, a ranking system using only the given probabilities was developed. Because human data were not available, classification derived from carcinogenic evidence in humans was not used (i.e. Group A or Group B1 WOE categories).

A QSAR carcinogenicity predicted probability ≤0.3 was designated as a noncarcinogen, and probability ≥ 0.7 was designated as a carcinogen (Accelrys, 2006). The range between 0.3 and 0.7 was considered the "indeterminate" zone. A substance with a carcinogenicity probability ≤0.3 was classified a category "E" carcinogen (noncarcinogenic) and not assigned a score to be used in the overall RQ/TES scoring. A substance with a carcinogenicity WOE probability $0.3< x <0.7$ was classified as a "C" carcinogen (possible human carcinogen) and assigned a score in the lowest potency group of 100 in accordance with EPA's

RQ protocol. A substance with a WOE carcinogenicity probability >0.7 was classified as a “B2” carcinogen (probable human carcinogen with no human evidence) and assigned a score of 10 in accordance with the EPA protocol.

Acute Toxicity

Acute toxicity was assessed using the QSAR rat oral LD_{50} model available in the computational software. The acute toxicity QSAR model consisted of 19 models in the rat oral LD_{50} module (Accelrys, 2006). The cross-validated accuracy of the 19 rat oral LD_{50} models is shown in Table 2. The rat oral LD_{50} model was based on experimental values from 4,000 substances from the Registry of Toxic Effects of Chemical Substances (RTECS). Only exposure times ranging from 0.5 to 14 hours were used.

Table 2. Cross-validated accuracy of the 19 acute toxicity models[a]

Chemical Class	Number of Compounds	% of Compounds predicted within a factor of				95% of Compounds predicted within a factor of
		2	3	4	5	
Organophosphates (P=0)	230	48	67	80	86	9
Organophosphates (P=5)	285	58	81	90	95	5
Carbamates	205	63	84	91	96	5
Heteroaromatics	429	63	83	92	97	5
Multiple Benzenes	367	70	85	92	95	5
Fused Benzenes	75	84	100			3
Single Benzenes (Subst =1)	196	80	96	99	100	3
Single Benzenes (Subst =2)	274	76	93	98	100	3
Single Benzenes (Subst =3)	162	80	92	97	100	4
Single Benzenes (Subst >3)	101	74	92	99	100	4
Alicyclics	361	65	85	93	97	4
Acyclic Amines	225	68	87	93	96	4
Acyclic Halo/Hydro-carbons	63	73	88	98	100	4
Acyclic Acids/Esters	138	67	89	98	100	3
Acyclic Alcohols	74	90	98	100		3
Acyclic Carbonyls	60	81	94	100		3
Acyclic Ethers	47	93	100			2
Acyclic C,O,H Miscellaneous	108	90	100			2
Acyclic (Others)	224	59	81	89	93	6

[a](Accelrys, 2006)

A TES value was assigned for a given rat oral LD_{50} according to the Oral Mammalian Toxicity Scale used to assign RQs for acute toxicity shown in Table 3.

Table 3. Oral Mammalian Toxicity Scale[a]

MAMMALIAN TOXICITY (ORAL)	TES
100 mg/kg ≤ LD_{50} < 500 mg/kg	5000
10 mg/kg ≤ LD_{50} < 100 mg/kg	1000
1 mg/kg ≤ LD_{50} < 10 mg/kg	100
0.1 mg/kg ≤ LD_{50} < 1 mg/kg	10
LD_{50} < 0.1 mg/kg	1

[a](ATSDR, 2011b)

Chronic Toxicity

EPA derived chronic toxicity scores from two primary attributes of each substance: the Minimum Effective Dose (MED) and the type of effect (EPA, 2011). It was important to use both attributes because the toxicity of a substance is a function of both its efficacy and the affect elicited (De Rosa, Stara, & Durkin, 1985). When an MED was based on animal data, a human-equivalent dose was derived using Formula 2.

Formula 2. Human-equivalent dosing formula

$$animal\ dose \left(\frac{\frac{mg}{kg}}{d}\right) x \left(\frac{animal\ weight}{70kg}\right)^{1/3} x\ 70kg$$

In order to assign an RQ for a substance, EPA conventionally multiplies a substance's dose rating (RV_d) developed from the MED with the dose rating developed from the severity of effect (RV_e). Both the RV_d and RV_s range from 1 to 10 with 10 being the most severe effect or highest dose. Thus, the chronic scores ranged from 1 to 100. However, the QSAR model was not able to predict an MED or the type of effect.

The chronic toxicity value predicted by the QSAR model for each substance was a rat oral chronic LOAEL. The rat oral chronic LOAEL QSAR model was based on experimental data from 393 substances using data from EPA, National Cancer Institute, NTP Technical Reports, FDA New Drug Applications, and citations from open literature (Accelrys, 2006). The cross-

validated accuracy of the five rat oral chronic LOAEL QSAR models is shown in Table 4. Unlike traditional LOAELS, the rat oral chronic LOAELs predicted by the QSAR model were not specific for any particular endpoint effect. Therefore, a novel approach was developed to score chronic toxicity.

Table 4. Cross-validated accuracy of the 5 chronic toxicity models[a]

Chemical Class	Number of Compounds	% of Compounds predicted within a factor of				95% of Compounds predicted within a factor of
		2	3	4	5	
Single Benzenes	130	66	88	94	98	5
Multiple Benzenes	83	70	92	96	97	4
Heteroaromatics	69	78	92	98	100	4
Alicyclics	39	94	100			3
Acyclics	73	73	92	97	100	4

[a](Accelrys, 2006)

In keeping with the approach of EPA, the predicted animal chronic LOAEL were used in lieu of the animal MED values. Animal chronic LOAELs were converted to human chronic LOAELs using Formula 1. A substance-specific rating value (RV_d) was determined by substituting the predicted human chronic LOAELs for MEDs in the chronic toxicity conversion scale using scientific judgment (EPA, 2011). Rating values ranged from 1-10.

Conventionally, a substance's RV_d is multiplied by a substance's rating value based on effect (RV_e) ranging from 1-10. However, a specific effect was not identified by the QSAR model. A developmental toxicity QSAR model that predicted a substance's probability of causing developmental toxicity was available and acceptable as a surrogate for evaluating chronic toxicity (ATSDR, 2011a) Consequently, developmental toxicity was used as a surrogate for level of effect (RV_e). The cross-validated accuracy of the three developmental QSAR models is shown in Table 4. The probability of potential developmental toxicity was multiplied by 10 and rounded to the nearest whole number to obtain a number ranging from 1-10 in accordance with the rating values for toxic effects.

Table 5. Cross-validated accuracy of the 3 developmental toxicity models[a]

Chemical Class	Number of Compounds	Specificity (%)	Sensitivity (%)	Indeterminate (%)
Aliphatics	87	88.6	88.6	2.5
Carboaromatics	95	97.4	87.0	2.2
Heteroaromatics	91	86.0	86.1	2.5

[a](Accelrys, 2006)

Substance-specific RV_ds and RV_es were multiplied to get a composite score. The possible range of composite scores was 1-100. The resulting composite scores were assigned corresponding TES scores according to the five tiers of scores in Table 6.

Table 6. Chronic Toxicity Scale[a]

COMPOSITE SCORE	TES
1-5	5000
6-20	1000
21-40	100
41-80	10
81-100	1

[a](Accelrys, 2006)

Aquatic Toxicity

Aquatic toxicity was assessed using the QSAR fathead minnow LC_{50} model available in the QSAR model. The aquatic toxicity QSAR model consisted of eight fathead minnow LC_{50} models developed from 444 studies (Accelrys, 2006). The cross-validated data were derived from open literature flow-through LC_{50} bioassays and five volumes on fathead minnow LC_{50} model developed by the Center for Lake Superior Environmental Studies. The cross-validate accuracy of the 8 fathead minnow models is show in Table 7.

Table 7. Cross-validated accuracy of 8 aquatic toxicity models[a]

Chemical Class	Number of Compounds	% of Compounds predicted within a factor of				95% of Compounds predicted within a factor of
		2	3	4	5	
Acyclic (halo/hydrocarbon)	33	81	100			3
Acyclic (alcohols)	38	83	100			3
Acyclic (miscellaneous)	92	64	86	96	100	4
Alicyclics	59	89	98	100		3
Multiple/Fused Benzenes	43	78	90	92	100	5
Single Benzenes (Subst= 1)	38	100				2
Single Benzenes (Subst= 2)	88	76	92	100		4
Single Benzenes (Subst=3)	53	78	94	96	98	4

[a](Accelrys, 2006)

A QSAR TES value was assigned for a given LC_{50} according to EPA's Aquatic Toxicity Scale used to assign RQs for aquatic toxicity shown in Table 8.

Table 8. Aquatic Toxicity Scale[a]

AQUATIC TOXICITY	TES
100 mg/l ≤ LC_{50} < 500 mg/l	5000
10 mg/l ≤ LC_{50} < 100 mg/l	1000
1 mg/l ≤ LC_{50} < 10 mg/l	100
0.1 mg/l ≤ LC_{50} < 1 mg/l	10
LC_{50} < 0.1 mg/l	1

[a](ATSDR, 2011b)

Biodegradability

EPA raised a substance's RQ by one tier if biodegradation, hydrolysis, and/or photolysis (BHP) resulted in degradation when the substance was released into the environment (EPA, 2011). An aerobic biodegradability model in the QSAR software was used to estimate the probability of environmental degradation of substances. The aerobic biodegradability QSAR model consisted of four structurally based sub-models (Accelrys, 2006). The cross-validated accuracy of these models is shown in Table 9. All data were determined

according to a Japanese Ministry of International Trade and Industry (MITI) 1 test protocol on 894 compounds, as cited in (Accelrys, 2006).

Table 9. Cross-validated accuracy of 4 aerobic biodegradability models[a]

Chemical Class	Number of Compounds	Validated leave-one-out accuracy %	Internal Accuracy %
Acyclics	317	96.1	97.7
Alicyclics	85	96.5	98.8
Single Benzenes	290	91.2	95.1
Multiple Benzenes and Heteroaromatics	160	93.1	98.1

[a](Accelrys, 2006)

The QSAR model predicted a probability of degradation for each substance, which was used as a surrogate for BHP activity. If a substance had a biodegradation probability of ≥ 0.7, then the overall RQ/TES for the substance was raised one tier. If a substance was highly volatile (boiling point ≤ 100° F), the RQ/TES of the parent compound was not raised one level. If the degradation products were more hazardous than the parent compound, then the parent compound was assigned the RQ/TES of the degradation products. For this study, an assumption was made that when the toxicity of a parent compound was not known, the relative toxicity of its metabolites was also unknown. Therefore, scoring was performed as if all metabolites were less harmful than the parent compound

Ignitability/Reactivity

EPA ignitability and reactivity scored were based on a substance's flash point, boiling point, and reactivity with water and self. The ignitability and reactivity values of substances were not able to be predicted computationally. The flash point and boiling point values for many substances were available through ChemIDplus Advanced (NIH, 2011) and the Hazardous Substances Data Bank (NLM, 2011). When a flash point and boiling point were both available, a TES was established using the scale in Table 10. When only a boiling point was

available, a substance was assigned a TES using the boiling point scale in Table 10 (BP <100°F, TES=100; BP>100°F, TES=1000).

Table 10. Ignitability Scale[a]

Ignitability	TES
FP 100°-140°F	5000
FP <100°, BP >100°F	1000
FP <100°, BP <100°F	100
Pyrophoric or self-ignitable	10
RQ of 1 not assigned based on ignitability/reactivity	

[a] (ATSDR, 2011b)

Each criterion was assigned a TES score based on the QSAR-predicted data. The lowest TES across all of a substance's criteria was assigned to each substance as the overall TES.

RESULTS

ANAYLSIS OF INTRA-CRITERIA AGREEMENT

ATSDR was mandated by CERCLA 1980 as amended by SARA 1986 to establish and revise a Substance Priority List of substances commonly found at NPL sites (ATSDR, 2008). The frequency of sites at which a substance was found, a substance's toxicity, and the potential for human exposure to a substance at NPL sites were used to rank substances on the SPL (Formula 1). In the past, RQs established by EPA were used by ATSDR to estimate a substance's toxicity. However, for 38 candidate SPL substances, toxicity information and RQs were not available. In addition, many substances ranked on the SPL did not have complete toxicity data for all toxicity endpoints considered (carcinogenicity, acute toxicity, chronic toxicity, aquatic toxicity, biodegradability, and ignitability/reactivity). Therefore, QSAR approaches were

used to computationally predict QSAR TES surrogates when the original data needed to rank substances on the ATSDR SPL were not available.

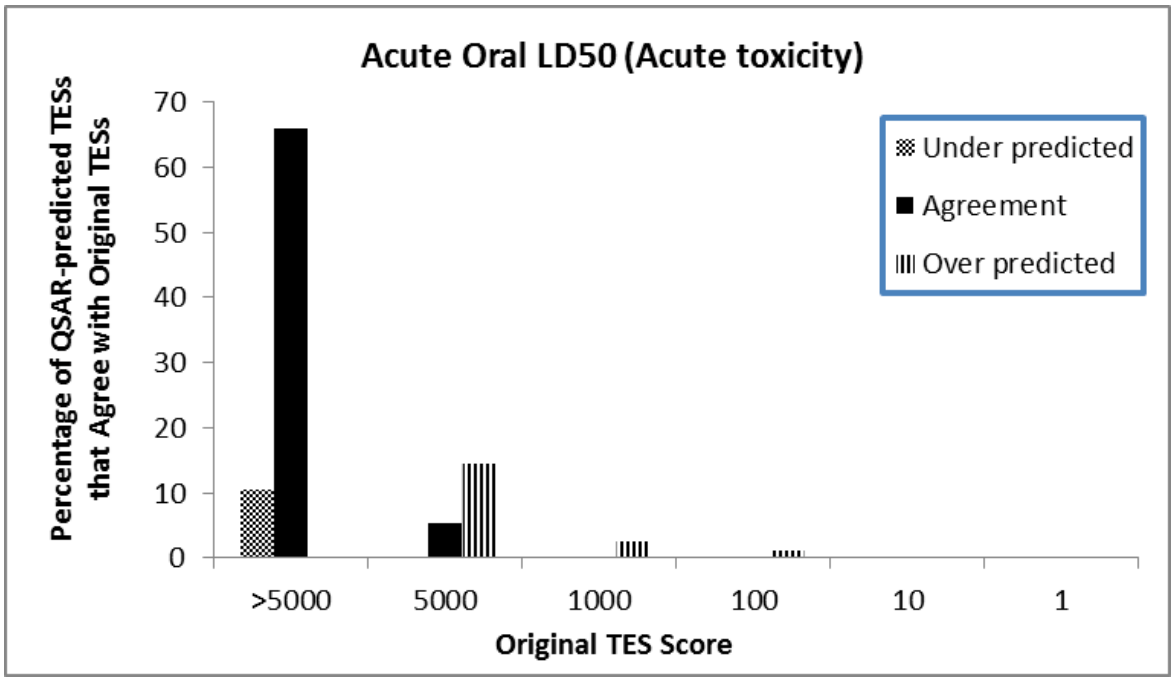
QSAR approaches were able to assess 293 of the 847 substances on the SPL. Of these substances, 191 were not able to be evaluated because they fell outside of TOPKAT's OPS. The remaining 102 substances were used this study

Acute Toxicity

Original and predicted scores for acute toxicity had a very high agreement (71%) (Figure 1 and Table 11). The QSAR model for acute toxicity was developed with data extracted from RTECS (Accelrys, 2006). RTECS listed the most toxic value when multiple values existed. Thus, acute toxicity QSAR model values were expected to be more conservative and overestimate toxicity, which corresponded to an underestimated (i.e. lower) toxicity score. Roughly 10% of the substances in the highest tier of scores were under predicted, and thus estimated to be more acutely toxic than shown in experimental studies (Figure 1).

Roughly 15% of substance TESs in the 5000 tier were over predicted by the QSAR model and were predicted to be less toxic than shown in experimental studies. It was impossible for substances assigned an original score of >5000 (the highest tier) to have an overestimated QSAR-predicted TES because no scores were assigned greater than the highest tier.

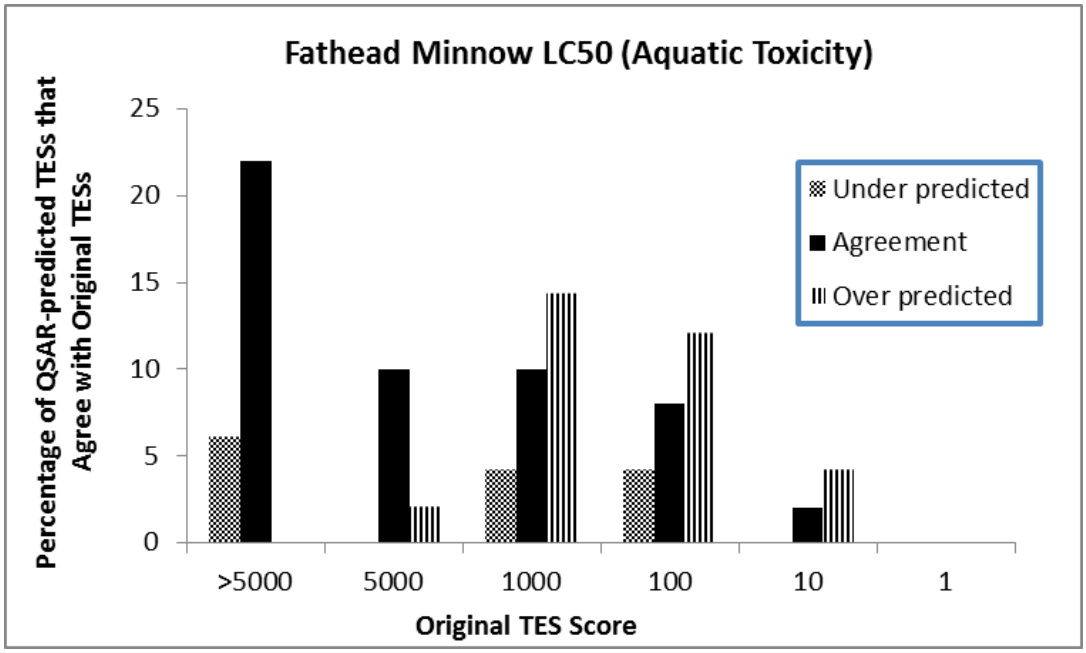
Figure 1. Analysis of agreement between original and QSAR-model predicted TESs for acute toxicity



Aquatic Toxicity

Likewise, aquatic toxicity was well predicted by the QSAR model (53% agreement) and had a moderate availability of original data (Figure 2 and Table 11). TESs predicted by the aquatic toxicity QSAR model that were not in agreement with original scores tended to be over predicted, implying an underestimate of the true toxicity. This was more prevalent in the lower tiers of scores.

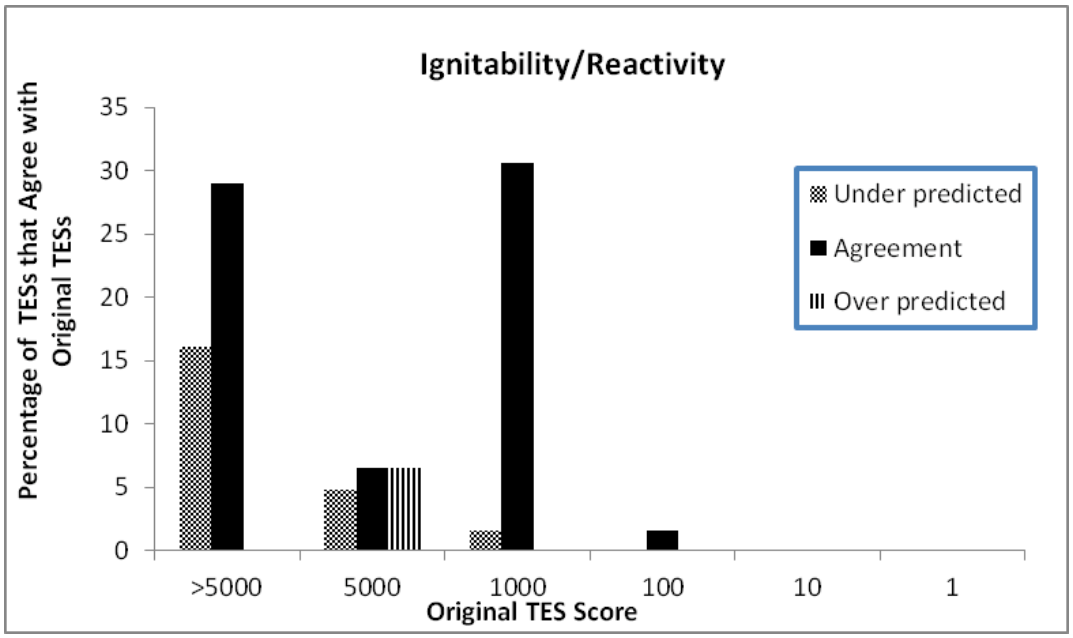
Figure 2. Analysis of agreement between original and QSAR-model predicted TESs for aquatic toxicity



Ignitability

Ignitability had strong agreement (69%) between original TESs and TESs calculated for this study (Figure 3 and Table 11). However, ignitability was not computationally predicted. TESs calculated for ignitability were based on experimental values obtained from peer-reviewed sources.

Figure 3. Analysis of ignitability/reactivity TESs



Biodegradation

BHP had a high agreement between original and predicted TESs (67%) (Table 11). BHP was not scored in tiers like the other criteria. Rather, it was scored based on a probability predicted by QSAR as "yes" if it was likely to degrade in the environment ($p \geq 0.7$) or "no" if it was not likely to degrade in the environment ($p < 0.7$). Thus, the percent agreement between original and predicted scores was high though little original data were available.

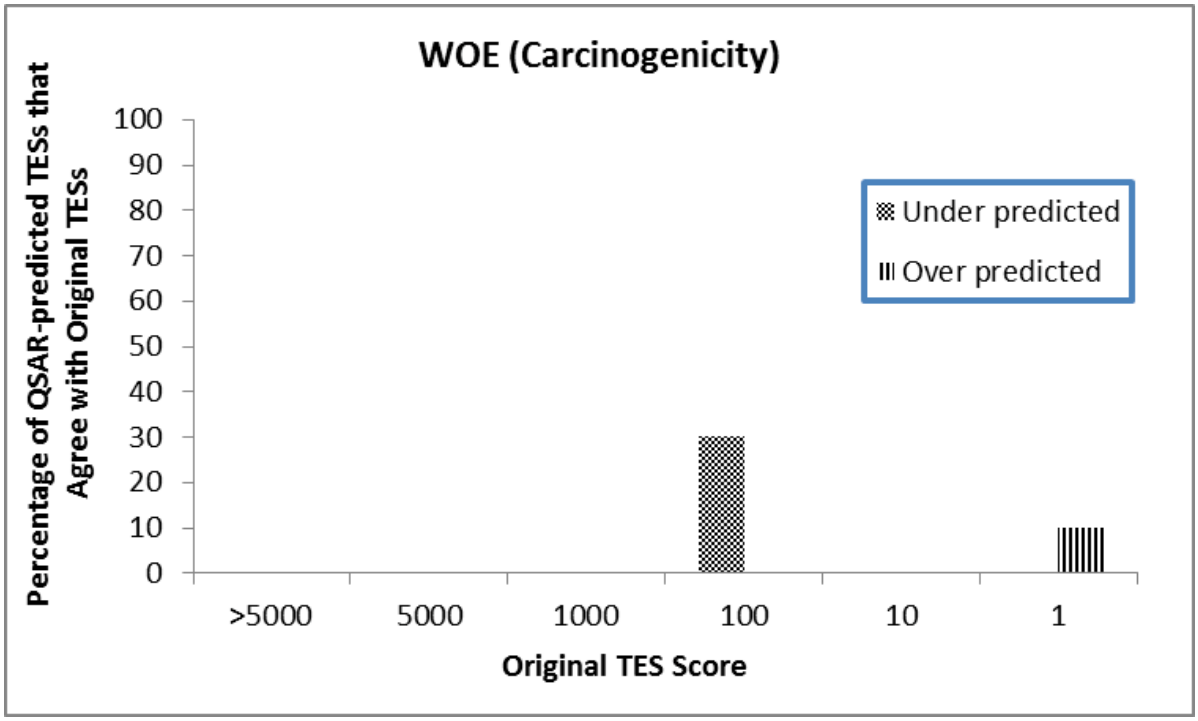
Carcinogenicity

Carcinogenicity QSAR-predicted TESs were not predicted accurately in this study (0%) (Figure 4 and Table 11). A number of studies have shown that carcinogenicity is not accurately predicted by existing computational models (Cronin, et al., 2003; Prival, 2001; Richard, 1998). However, all QSAR-predicted carcinogenicity TESs were within 1-tier of the original scores. The carcinogenicity QSAR model was more conservative than the available original data, but it did predict the probability of carcinogenicity within a score factor of 10.

If a substance was predicted to have a probability of carcinogenicity ≤0.3, then it was not assigned a TES for carcinogenicity (Accelrys, 2006). Only 40% (n=4) of the 10 substances that had original data had a probability of carcinogenicity >0.3 and were assigned a corresponding TES. The QSAR model generally predicted toxicity to be more severe than experimental studies supported (Figure 4).

Carcinogenicity/weight of evidence (WOE) had such low agreement that an additional overall scoring approach was performed using all criteria except WOE. Table 12 shows the TES agreement and percent of TESs within ± 1 tier of the original TES among all scoring methods. Excluding WOE from the scoring approach nearly doubled the final TES agreement between original and predicted values.

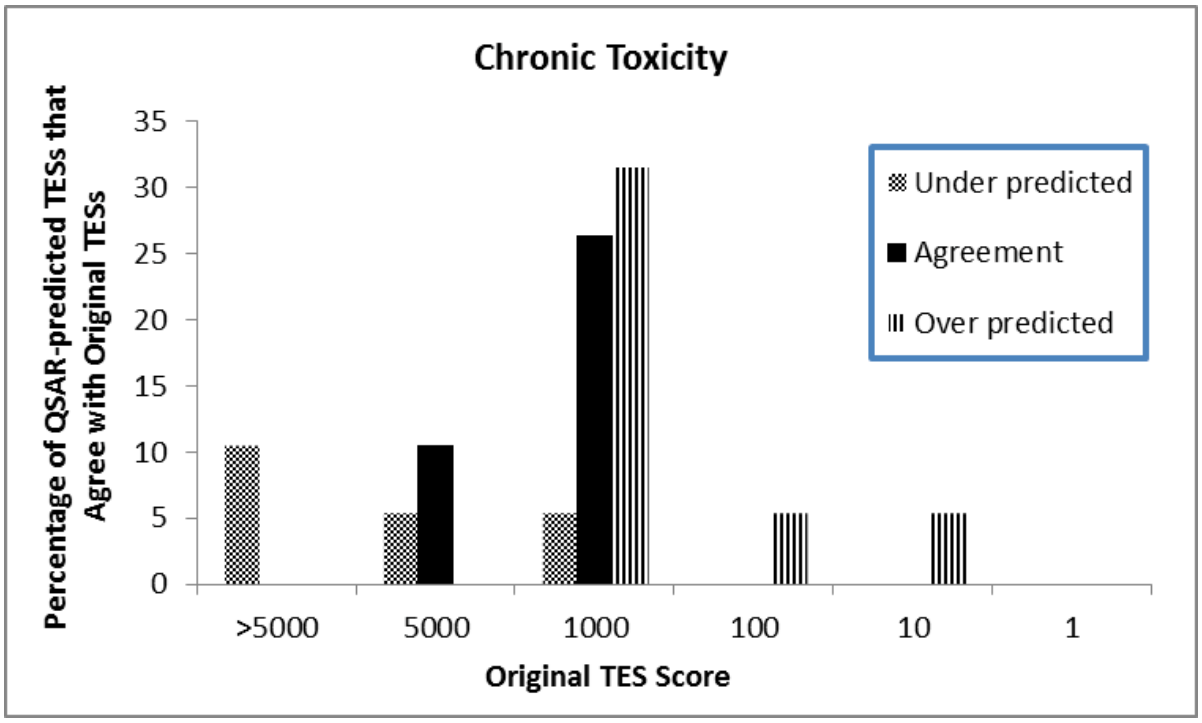
Figure 4. Analysis of agreement between original and QSAR-model predicted TESs for carcinogenicity



Chronic Toxicity

Chronic toxicity was poorly predicted in this study (Figure 5 and Table 11). A novel approach substituting predicted probabilities of developmental toxicity in place of values for a specific effect was used to score chronic toxicity values. This approach nearly doubled the agreement observed within the chronic toxicity criteria. This approach tended to over predict TESs and under estimate toxicity (Figure 5 and Figure 6).

Figure 5. Analysis of agreement between original and QSAR-model predicted TESs for chronic toxicity

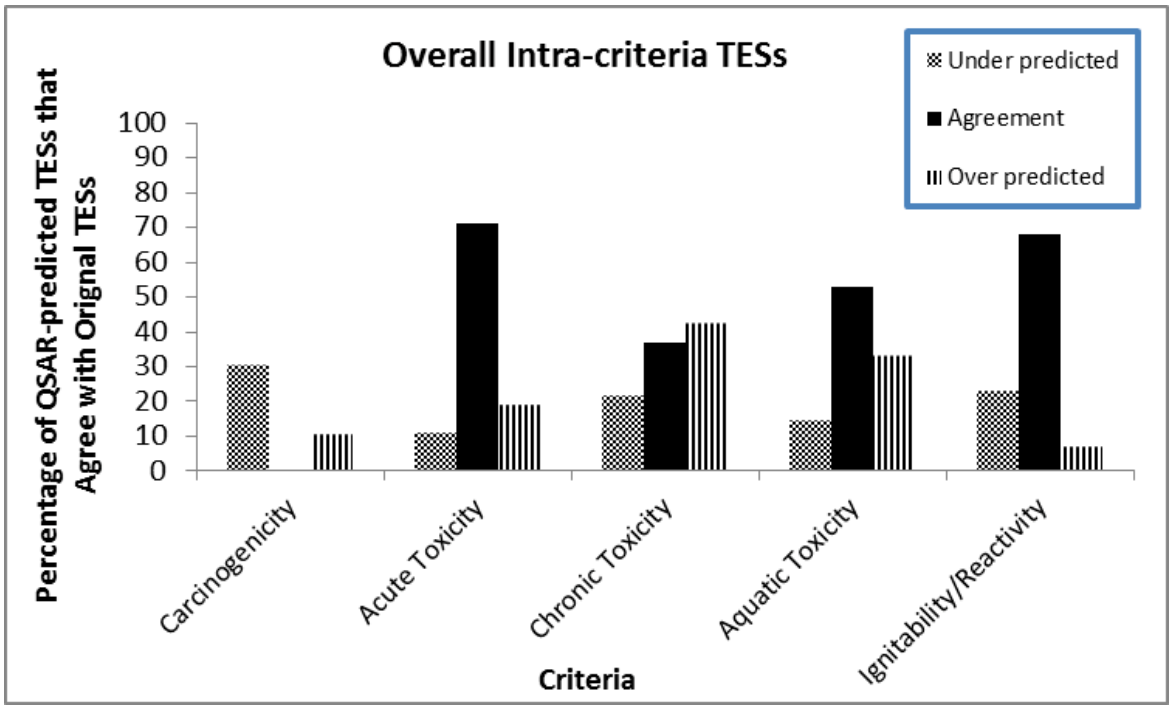


The agreement between original and predicted QSAR TESs is shown in Table 11 and Figure 6. Percentage agreement was only calculated for criteria that had experimental values, thus the denominators varied across criteria.

Table 11. Analysis of Original and Predicted TESs

Analysis	Weight of Evidence	Acute Toxicity	Chronic Toxicity	Aquatic Toxicity	Biodegradation/ Hydrolysis/ Photolysis	Ignitability
All chemicals in TOPKAT OPS	0% (0/10)	71% (54/76)	37% (7/19)	53% (26/49)	67% (6/9)	69% (43/62)

Figure 6. Analysis of overall intra-criteria QSAR-model predictions



ANALYSIS OF SCORING METHODS

Original and predicted TESs were compared and percentage agreements analyzed for overall scoring methods (Table 12). The percentage of TESs within 1 tier of the original score was determined in order to assess prediction accuracy. An overall analysis utilizing all QSAR-predicted scores had very poor original/predicted score agreement (15%), but a number of other promising methods were identified.

EPA's original substance RQ scoring approach used only available experimental data. Most substances did not have sufficient data to develop an original TES for all the criteria (ATSDR, 2011a). Agreement between original and predicted scores generally increased with increased availability of original data (Table 11). Therefore, overall TES agreement was not expected to be high among substances that lacked original data.

Analyses were performed to how the agreement between original and predicted overall TESs would change if parameters without original scores were dropped from the scoring method. Results are shown in Table 12. This approach produced the strongest TES agreement of 57%. In addition, 89% of all final TESs were within 1 tier of original TESs. Such results were not unexpected in lieu of the fact that the highest agreement within criteria was observed for those substances with the most original data. All other scoring methods incorporate predicted scores in the calculation of overall TESs even though one or more of the criteria lack original data.

QSAR-predicted carcinogenicity scores had such low agreement with experimental scores that these predicted scores were dropped from the scoring procedure in order to increase overall TES agreement. The analysis excluding WOE had very low TES agreement (29%) (Table 12). However 74 % of TESs in the analysis excluding WOE fell within 1 tier of the original score (Table 12). Additional computational models are needed that include a mechanistic approach to predicting carcinogenicity

Table 12. TES Final Score Agreement

Category	TES agreement	TESs ± 1 tier away from original score
Overall	15% (10/102)	56% (57/102)
Overall without WOE	29% (30/102)	74% (75/102)
Only using parameters with experimental data	57% (45/79)	89% (70/79)

DISCUSSION

In response to the congressional mandate in The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) of 1980, as amended by the Superfund Amendments and Reauthorization Act (SARA) of 1986, the EPA and ATSDR prepare and revise the Substance Priority List (SPL) to prioritize substances for Toxicological Profile development (ATSDR, 2008). Substances are ranked using three criteria according to the algorithm depicted in Formula 1. RQs developed by EPA are used by ATSDR as estimators of toxicity for the toxicity component of this algorithm.

RQs are derived from substance-specific experimental data including intrinsic physicochemical (ignitability/reactivity), toxicological properties (aquatic toxicity, acute mammalian toxicity, chronic toxicity, potential carcinogenicity), and biodegradation, hydrolysis, and photolysis (BHP). When RQs are not available, the ATSDR uses an equivalent approach to develop Toxicity/Environmental Scores (TESs). However, sufficient original data are not available to develop a TES for the potentially hazardous chemicals of concern. Quantitative Structure-Activity Relationship (QSAR) approaches can be used to computationally predict the physicochemical, toxicological and biodegradability properties needed to calculate surrogate TESs. QSAR methods were used to computationally predict TES surrogates when the experimental data needed to rank substances on the ATSDR SPL were not available. This study was limited by the number of substances available from the NPL listing activity and by the substance structures that could be reliably predicted by the QSAR software. One hundred and two substances were used this study.

Analysis of agreement within criteria (carcinogenicity, acute toxicity, chronic toxicity, aquatic toxicity, biodegradability, and ignitability) showed that agreement between original and predicted scores generally increased with increased availability of original data (Table 11). Toxicity estimates and TESs were inversely related. A substance with very high toxicity received a very low TES and vice versa. Thus, if a TES was overestimated, then the toxicity was predicted to be less severe than shown in experimental studies.

Agreement between original and predicted TESs is expected to be much higher when original data are available to compare with predicted data. Thus, as more experimental data become available, the agreements observed in criteria and overall are expected to increase.

Acute toxicity had the highest agreement (71%) between original and predicted TESs among criteria evaluated (Figure 1, Figure 6, and Table 11). This parameter had the most original data values, which is most likely a factor influencing the increased percentage agreement. Rat oral LD_{50} studies are less resource-intensive, less complex, and generally more easily interpreted than chronic toxicity or carcinogenic studies. This parameter may be better represented in the computational software database (and thus better predicted) because of the wealth of information available in scientific literature on LD_{50s}. Thus, rat oral LD_{50} toxicity predictions would be expected to be more accurate than predictions derived from other criteria.

Likewise, aquatic toxicity was well-predicted by the QSAR model (53% agreement) and had a moderate amount of original data (Figure 2, Figure 6, and Table 11). If this parameter had more available original data, agreement between original and predicted TESs would predictably be higher. Fathead minnow LC_{50} studies are relatively less resource-intensive than other studies and are thus well-represented in the QSAR database, thereby leading to higher prediction accuracy.

The Ignitability and Reactivity scoring approach had high agreement (69%) between original TESs and TESs calculated for this study (Figure 3, Figure 6, and Table 11). The scores were not computationally predicted. They were gathered from the peer-reviewed literature. The approach used to score this criterion and the high agreement showed that EPA ignitability/reactivity scoring approach could be applied to a wide variety of data sources, though it does not have direct implications for QSAR methods.

BHP was not scored in tiers like the other criteria. Instead it was scored based on a probability predicted by QSAR as “yes” if it was likely to degrade in the environment ($p \geq 0.7$) or “no” if it was not likely to degrade in the environment

(*p*<0.7*)*. Original s and predicted TESs in the BHP criteria are expected to have a higher percent agreement compared with other criteria because agreement is based on two available values (i.e. yes or no) instead of five as in the 5-tiered approach used for the other criteria. The BHP criteria only had eight original scores and would have been expected, if following the trend of the other criteria, to have a low agreement. However, the agreement between original and predicted scores is 67% (Table 11 and Figure 6).

Carcinogenicity/WOE was not predicted well in this study (Figure 4, Figure 6, and Table 11). The carcinogenic nature of a substance encompasses not only a substance's structure, but also metabolism and potential mechanisms of action (MOA). Though two substances may be structurally similar, the metabolism and MOA of these two substances may be very different (Prival, 2001). Computational software does not account for metabolism and MOAs and, therefore, poorly predicts carcinogenicity (Prival, 2001). Additional computational models are needed that include a mechanistic approach to predicting carcinogenicity.

Similarly, chronic toxicity is poorly predicted in this study and previous studies (Rupp, et al., 2010) (Figure 5, Figure 6, and Table 11). Original LOAEL scores are based on specific toxicity endpoints or biological activities related to a specific MOA. TOPKAT software predictions are based on combined toxicity endpoints (Venkatapathy, Moudgal, & Bruce, 2004). The original and predicted scores are indicative of different endpoints. This may explain the low agreement observed between original and predicted scores. A novel approach substituting predicted probabilities of developmental toxicity in place of values for a specific effect was used to score chronic toxicity values. Developmental toxicity probabilities are representative of the severity of developmental toxicity, which may or may not be indicative of overall toxicity. Additional studies exploring this novel approach of chronic toxicity using more substances are needed.

The highest agreement between overall TESs resulted from the scoring approach that selected final TESs only from criteria that had original TESs (Table 12). This is expected because the highest percent agreement within criteria is

observed for substances with the most original data. All other scoring approaches incorporate predicted scores in overall TESs, even when criteria lack original data.

The original overall TESs were not reflective of the criteria that lacked original data. Thus, though the predicted scores may be accurate, incorporating them into the overall scoring influences the overall TESs in a manner not observed in the original and additional scoring approaches. Therefore, a lower agreement is observed for these approaches. If original data existed for every criterion for every substance, then additional approaches may have had very high agreements.

This study shows that QSAR methods can be applied to data gaps and used to develop surrogate values to provide a comprehensive identification of a substance's toxicity. There exists an increasingly great need for toxicity endpoint data as more and more substances are introduced and the need for knowledge pertaining to substance hazards increases. Using the QSAR methods discussed in this study, these data gaps can be filled in less resource-intensive manner than required for experimental studies. Candidate substances for toxicological profile development can be ranked based on their toxicity despite existing experimental data gaps. This allows dissemination of pertinent health information and allows regulatory decisions to be made much quicker than if waiting for a lengthy experimental chronic or carcinogenicity study. In the future, multiple QSAR methods will be applied to develop the most accurate system for predicting endpoint toxicity and TES values, especially for substances that fell outside of TOPKAT's OPS.

REFERENCES

Accelrys. (2006). Toxicity prediction by komputer assisted technology (TOPKAT) User guide (Version 6.2). Burlington, MA.

ATSDR. (2008). FY 1996, Agency Profile and Annual Report: U.S. Department of Health and Human Services, Atlanta, GA.

ATSDR. (2011a). *2011 CERCLA priority list of hazardous substances that will be the subject ot toxicology profiles and support documents*.

ATSDR. (2011b). *EPA Reportably Quantity Methodology Used to Establish Toxicity/Environmental Scores for the 2011 Substance Priority List*. Retrieved from http://www.atsdr.cdc.gov/SPL/resources/ATSDR_2011_TES_Methodology.pdf.

Barratt, M. D. (1998). Integrating computer prediction systems with in vitro methods towards a better understanding of toxicology. *Toxicology Letters, 102–103*(0), 617-621. doi: 10.1016/s0378-4274(98)00266-5

Cronin, M. T., Walker, J. D., Jaworska, J. S., Comber, M. H., Watts, C. D., & Worth, A. P. (2003). Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. [Research Support, Non-U.S. Gov'tReview]. *Environ Health Perspect, 111*(10), 1376-1390.

De Rosa, C. T., Stara, J. F., & Durkin, P. R. (1985). Ranking chemicals based on chronic toxicity data. *Toxicol Ind Health, 1*(4), 177-191.

Demchuk, E., Ruiz, P., Chou, S., & Fowler, B. A. (2011). SAR/QSAR methods in public health practice. *Toxicol Appl Pharmacol, 254*(2), 192-197. doi: 10.1016/j.taap.2010.10.017

el-Masri, H. A., Mumtaz, M. M., Choudhary, G., Cibulas, W., & De Rosa, C. T. (2002). Applications of computational toxicology methods at the Agency for Toxic Substances and Disease Registry. *Int J Hyg Environ Health, 205*(1-2), 63-69.

EPA. (2011). Reportable Quantities, from http://www.epa.gov/superfund/policy/release/rq/index.htm#method

McKinney, J. D., Richard, A., Waller, C., Newman, M. C., & Gerberick, F. (2000). The Practice of Structure Activity Relationships (SAR) in Toxicology. *Toxicological Sciences, 56*(1), 8-17. doi: 10.1093/toxsci/56.1.8

Mumtaz, M. M., Knauf, L. A., Reisman, D. J., Peirano, W. B., DeRosa, C. T., Gombar, V. K., . . . et al. (1995). Assessment of effect levels of chemicals from quantitative structure-activity relationship (QSAR) models. I. Chronic lowest-observed-adverse-effect level (LOAEL). *Toxicol Lett, 79*(1-3), 131-143.

NIH. (2011). ChemIDplus Advanced. from National Library of Medicine, National Institute of Health <http://chem.sis.nlm.nih.gov/chemidplus/>

NLM. (2011). Hazardous Substances Data Bank (HSBD). from United States National Library of Medicine http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB

Prival, M. J. (2001). Evaluation of the TOPKAT system for predicting the carcinogenicity of chemicals. [Evaluation Studies]. *Environ Mol Mutagen, 37*(1), 55-69.

Richard, A. M. (1998). Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet? [Review]. *Mutat Res, 400*(1-2), 493-507.

Ruiz, P., Mumtaz, M., & Gombar, V. (2011). Assessing the toxic effects of ethylene glycol ethers using Quantitative Structure Toxicity Relationship models. *Toxicol Appl Pharmacol, 254*(2), 198-205. doi: 10.1016/j.taap.2010.10.024

Rupp, B., Appel, K. E., & Gundert-Remy, U. (2010). Chronic oral LOAEL prediction by using a commercially available computational QSAR tool. [Research Support, Non-U.S. Gov't]. *Arch Toxicol, 84*(9), 681-688. doi: 10.1007/s00204-010-0532-x

Venkatapathy, R., Moudgal, C. J., & Bruce, R. M. (2004). Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. [Research Support, U.S. Gov't, Non-P.H.S.]. *J Chem Inf Comput Sci, 44*(5), 1623-1629. doi: 10.1021/ci049903s