Towards guidance on how to characterize predictive uncertainty in QSAR regression models

— within the CADASTER project —

Ullrika Sahlin[1], Tom Aldenberg[2], Tomas Öberg[1]

[1]Linnaeus University, Sweden. [2]RIVM, Netherlands.

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

Linnæus University

CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Work package 4 - Integration of QSARs within hazard and risk assessment

Subtask 1 - Integration of QSAR models into a probabilistic risk assessment framework

Deliverable 1 - Application of QSAR models for probabilistic risk assessment

Deliverable 2 - Guidance on using QSAR models for probabilistic risk assessment

SEVENTH FRAMEWORK PROGRAMME

Outline

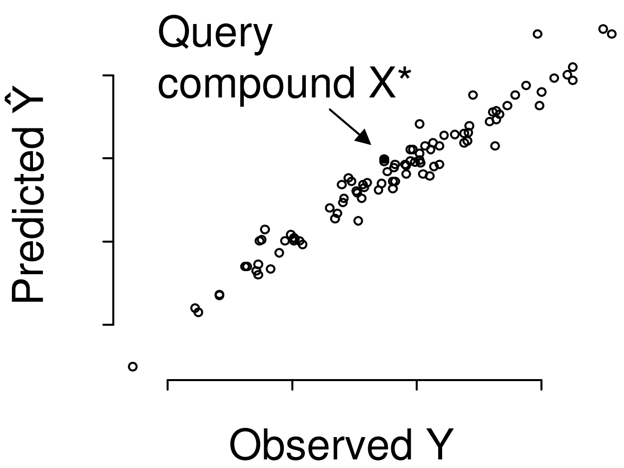
1. Predictive uncertainty
2. Compilation of methods to assess predictive uncertainty*
3. The methods will be evaluated with respect to
   3.1 Theoretical and statistical aspects
   3.2 The ability to reproduce prediction uncertainty empirically
   3.3 The intended use (e.g. easiness of implementation or perceived as acceptable by risk assessors).
4. Conclusions and future outlook

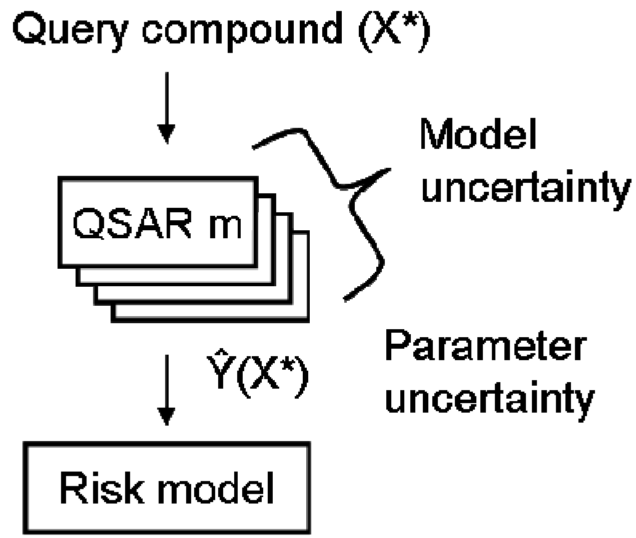
* Sahlin et al 2011. A Risk Assessment Perspective of Current Practice in Characterizing Uncertainties in QSAR Regression Predictions. DOI: 10.1002/minf.201000177

1. Predictive uncertainty – a risk assessment perspective

Parameter uncertainty – uncertainty in predicted values of query compound

Model uncertainty – uncertainty in using the QSAR to predict the query compound





1. Predictive uncertainty - different types of characterizations

Parameter uncertainty – Predictive distribution

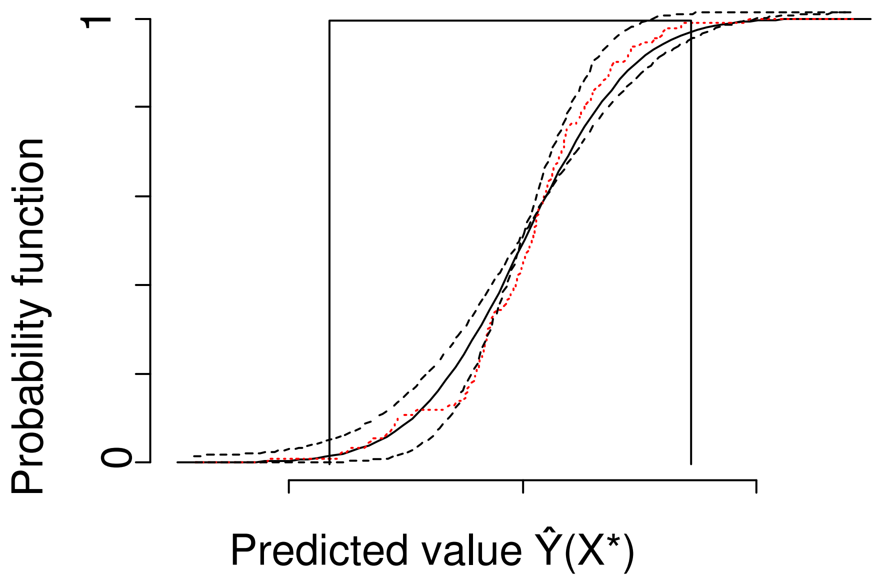
Parametric or empirical probability distribution

2-dimensional probability distribution

Interval (fuzzy number)

Combination of these –

probability box



1. Predictive uncertainty - different types of characterizations

Model uncertainty – Measures of Reliability

- Reliability may depend on the relation to the applicability domain
- Reliability measures for classification QSARs could be applicable on regression
- Reliability follows from assessment of predictive uncertainty – e.g. empirical coverage, number of compounds that fall inside prediction intervals of a given confidence level, distance to the predictive distribution, …

2. Compilation of methods to assess predictive uncertainty

Predictive distribution may be assessed

- from **estimates of predictive variance** (e.g. by sampling or re-sampling). Necessary when QSAR regression predict point estimates – e.g. least square regression, kNN
- directly as **probability distributions** – e.g. Bayesian linear regression, generalized linear model, density estimator
- based on experimental data – **expert judgement**

1. Predictive variance depend on the **applicability domain** (e.g. distance to AD: leverage, density of AD: DPRESS)
2. Consensus-modelling – there is **no best model**, predict by averaging models, predictive uncertainty from the variance over several models

3. Evaluation – theoretical and statistical aspects

In a decision theoretic framework:

- ”A predictive model minimizes the expected loss”
  - Required: Loss function and means to derive an expectation

- QSAR predictions are given as: conditional expectations, conditional densities or predictive posteriors.

3. Evaluation – theoretical and statistical aspects

- Predictive uncertainty can be assessed by approaches being
  1. Parametric (e.g. Bayesian)
  2. Empirical (e.g. sampling)

- Both Bayesian and empirical approaches generates predictive uncertainty with probabilistic interpretation.

3. Evaluation – theoretical and statistical aspects

Bayesian approach

E.g. Bayesian linear regression (BLR)

- Parametric model $Y|X \sim N(\beta X, \sigma^2)$
- Predictive distribution from likelihood and prior distributions on parameters β's and σ^2.
- Predictive variance is equal to $(1+X^*(X'X)^{-1}X^{*\prime})\sigma^2$

- BLR offers a straightforward assessment of predictive uncertainty, <u>but</u> QSAR data is ”small n large p”.

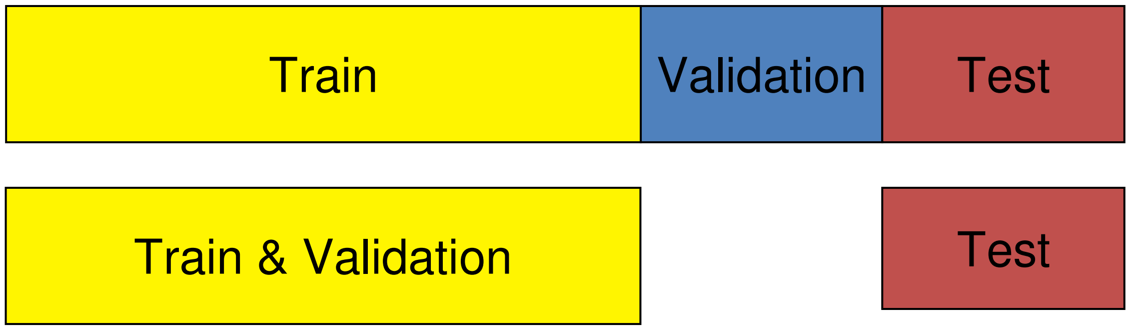
3. Evaluation – theoretical and statistical aspects

Empirical approach

Predictive variance $V(\hat{Y}(X^*))$ = PRESS/n or DPRESS/n

Predictive distribution – assigned by the assessor

Note: Cross-validation generate MEAN and STD for predictive variance



PRESS = PRedictive Error Sum of Squares
DPRESS = Distributed PRESS

3. Evaluation – theoretical and statistical aspects

Bayesian – comparison

<u>Pros:</u> Assess uncertainty directly based on data, and prior knowledge – theoretically underpinned. Can combine empirical data and expert judgement.

<u>Cons.</u> Difficult to implement in practise, requires understanding of difficult mathematical language. Difficulties in matching small data sets with many descriptors.

Empirical – comparison

<u>Pros:</u> Works with any type of underlying algorithm, Easy to calculate.

<u>Cons:</u> Sampling sensitive to the availability and choice of test set

For small data no external data set is available – rely on internal CV

3. Evaluation – the ability to reproduce predictive uncertainty empirically

CADASTER consensus modeling - ”there is no best model”

- Predictive QSARs generated by alternative algorithms using
  - best practice + methods to assess predictive uncertainty
- Evaluate
  - Predictivity (according to OECD principles)
  - Reliability (based on assessed predictive uncertainty)

Table 1. MP statistical parameters verified for the development of QSAR models for MP

Train	Prediction	Group	R^2	Q^2_{LOO}	Q^2_{BOOT}	$RMSE_{TR}$	$RMSE_{Ext}$	Q^2_{F1}	Q^2_{F3}	Q^2_{Ext}	R^2_{YS}
53	Prediction Ia	UI	83.34	79.31	77.82	33.95	51.69	76.09	61.37†	-	7.46
	15 EV-set				77.2		24.92	72.45	91.02		7.6
	Prediction (15 EV-set)	LNU	90.20	85.20		27.23	15.62			85.45	
	Prediction Ia	IDEA	86.10			31.03	51.71	76.10	61.30[2]	-	-
	15 EV-set						36.88	39.70	80.30	-	-
	Prediction Ia	HMGU	80	-	-	39	46	-	-	-	-
	15 EV-set						39				

3. Evaluation – the ability to reproduce predictive uncertainty empirically

Exercise – compare

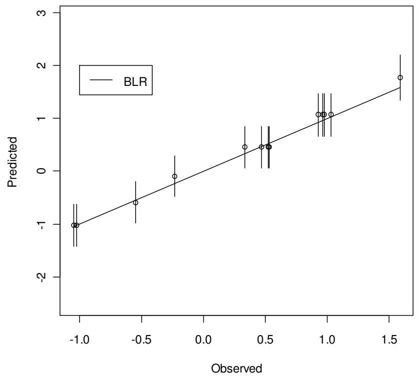
- Bayesian linear regression with model selection (BLR)
- Multivariate linear regression and sampling (MLR + PRESS, MLR + DPRESS)
- Bayesian lasso regression (blasso)
- Shrinkage regression and sampling (lasso + PRESS)
- Partial Least Square regression and sampling (PLS + PRESS)

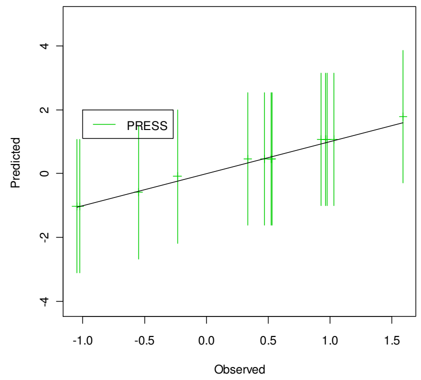
Here fitted to QSAR data on

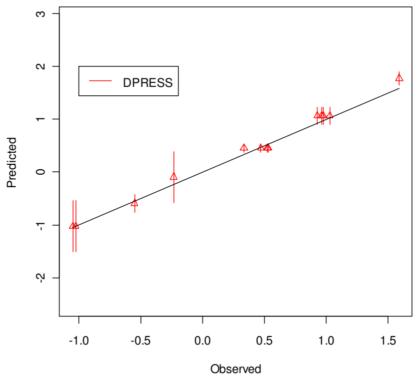
- Vapour pressure for PBDEs (property endpoint)
- Fish LC_{50} for benzotriazoles (effect endpoint)

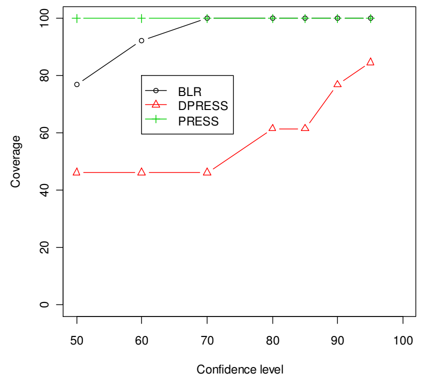
3. Evaluation – BLR vs PRESS vs DPRESS

Property endpoint



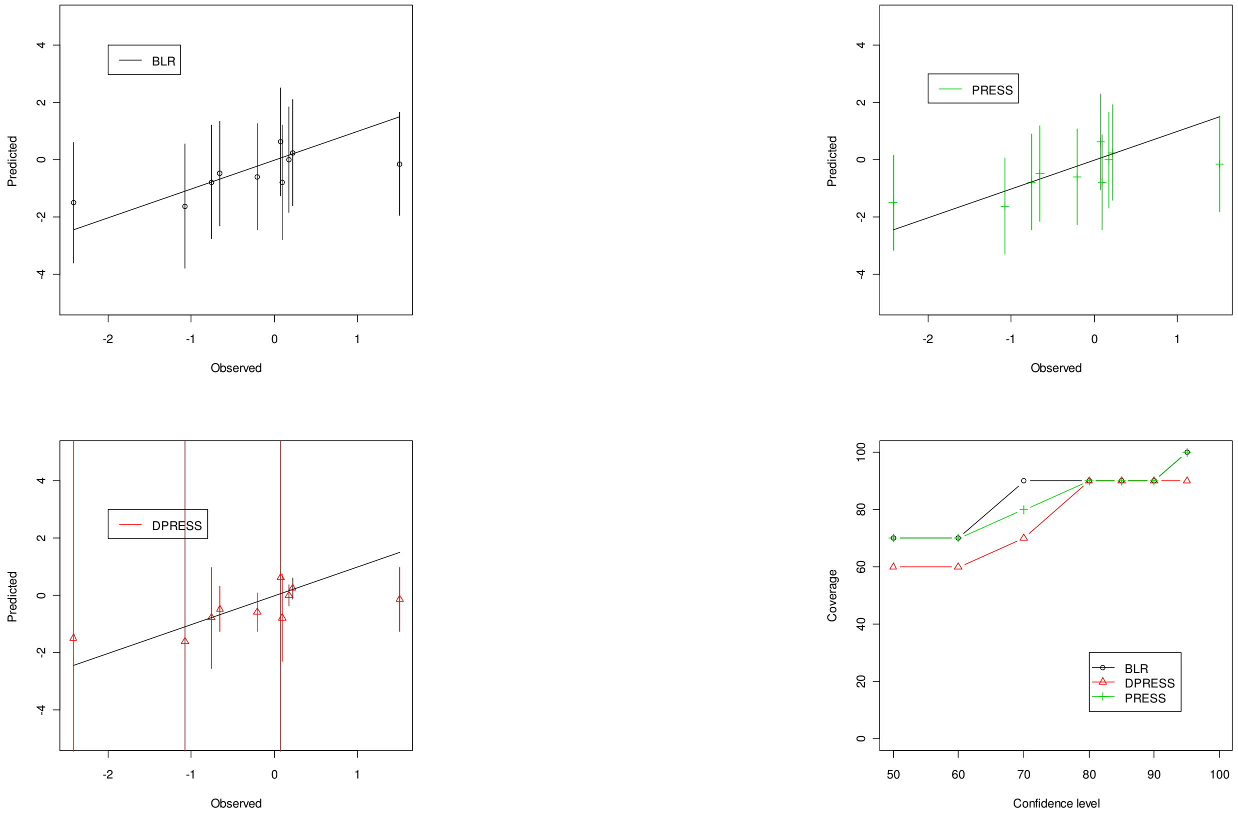






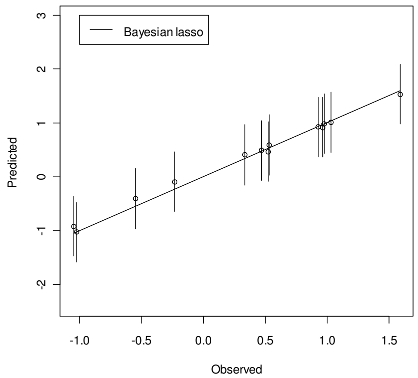
3. Evaluation – BLR vs PRESS vs DPRESS

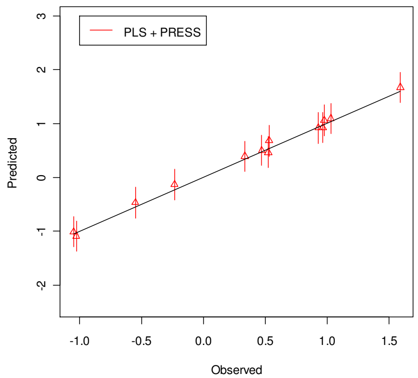
Effect endpoint

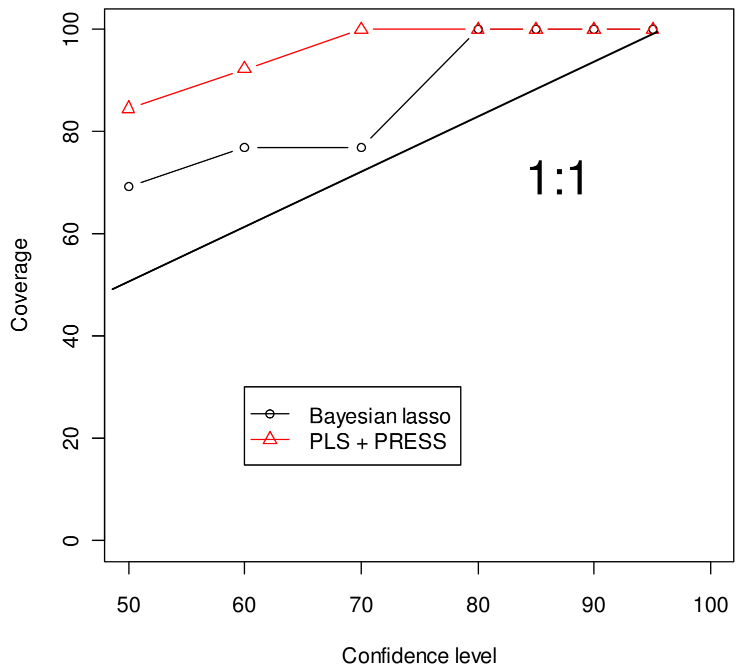


3. Evaluation – Bayesian lasso vs PLS

Property endpoint

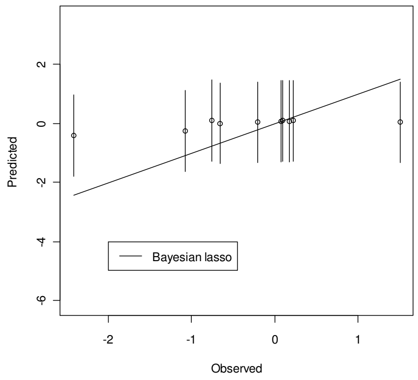


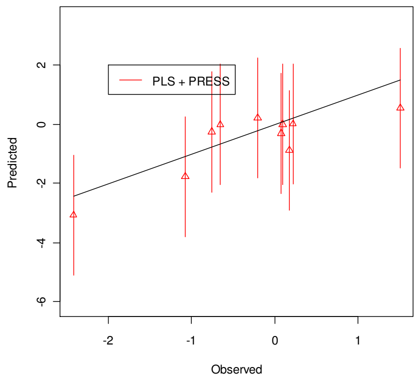


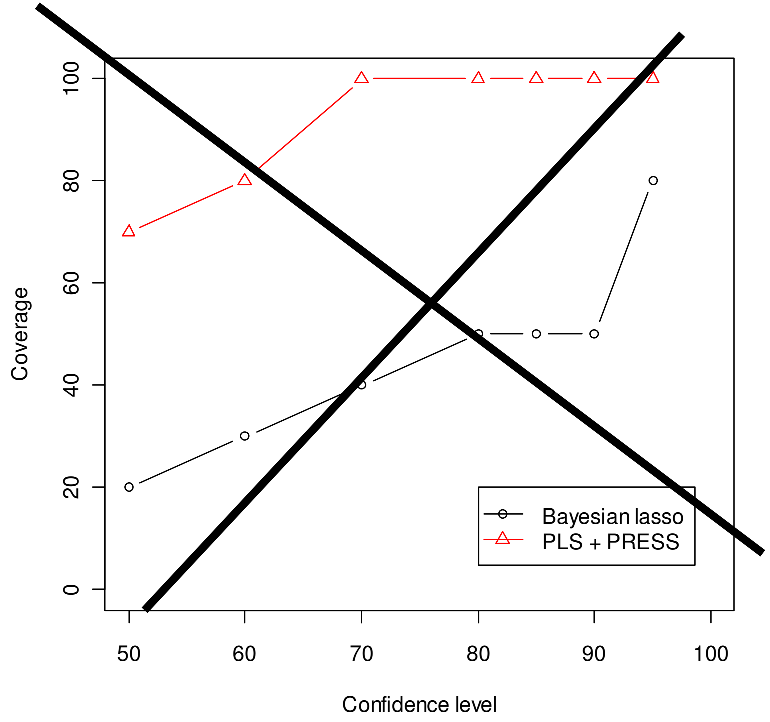


3. Evaluation – Bayesian lasso vs PLS

Effect endpoint

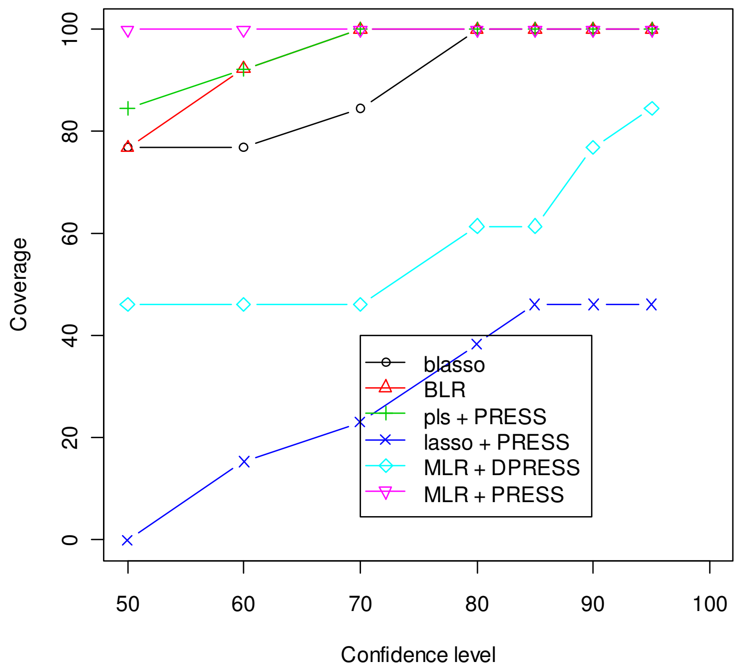




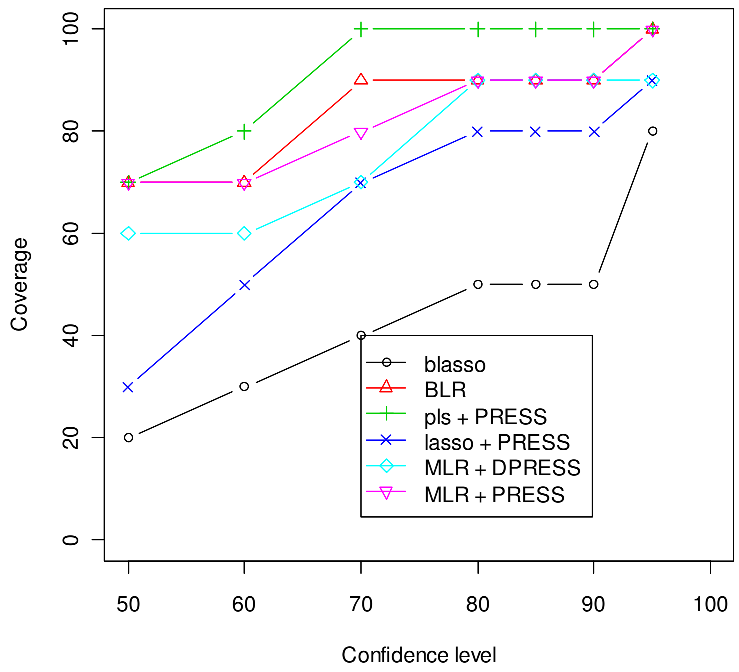


3. Evaluation – ongoing…

Property endpoint



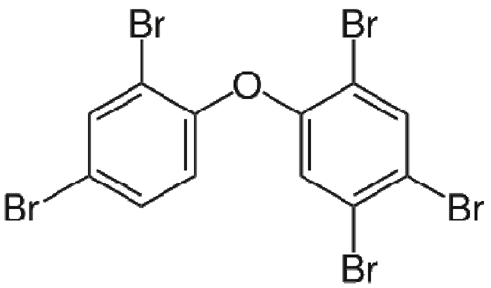
Effect endpoint



3. Evaluation – the intended use

- What kind of characterization of predictive uncertainty is needed for risk assessment or weight-of-evidence approaches?
- What measures of reliability are useful?
- Which methods for characterization are most appealing to end-users?
- When does it matter which methods to use?

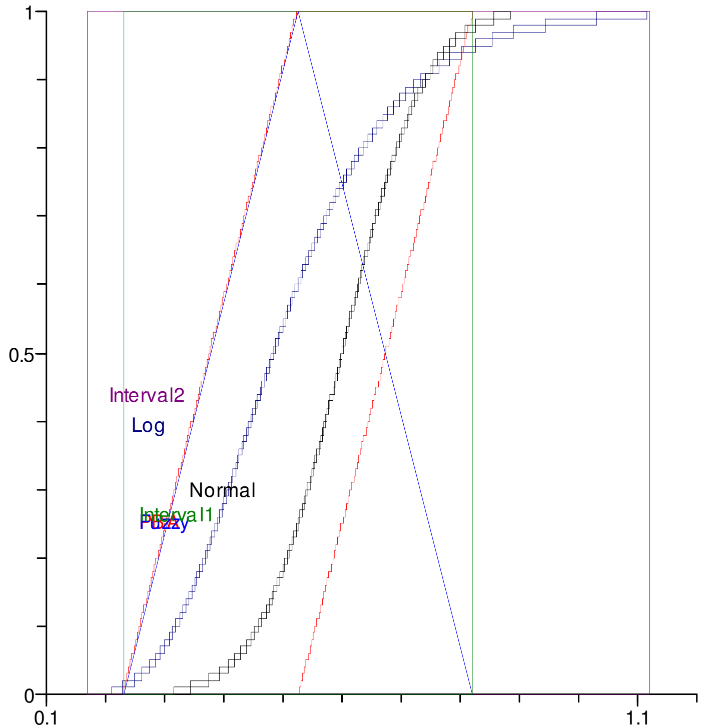
3. Evaluation – the intended use



Different approaches to describe uncertainty in risk models. Example HL for BDE-99

- Experimental uncertainty:
  - one lab N(0.60,0.11)
  - between labs 0.23-0.82
- QSAR (2*RMSE$_T$):
  - 0.51±0.34 (0.17-1.12)

Normal=N(0.6,0.11)
Interval1=[0.23,0.82]
Fuzzy=[0.23,0.525,0.82]
PBA=U([0.23,0.525],[0.525,0.82])
Log=L(0.51,0.17)
Interval2=[0.17,1.12]



4. Conclusions and future outlook

- The characterization of predictive uncertainty is not regulated for QSAR regressions
- Predictive uncertainty ask for probabilistic QSARs and statistical predictive inference
- Methods that assess predictive uncertainty needs to be evaluated together with algorithms of model building.
- How to characterize predictive uncertainty depend on the context and purpose of prediction
- Any recommended method(s) must be general enough to encompass a range of different model building approaches

Your input is needed. Please contact us with questions and suggestions!

More info: Ullrika.Sahlin@lnu.se

Rijksinstituut voor Volksgezondheid
en Milieu
Ministerie van Volksgezondheid,
Welzijn en Sport

Linnæus University