Regulatory Toxicology and Pharmacology 69 (2014) 443–450

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Developing scientific confidence in HTS-derived prediction models: Lessons learned from an endocrine case study

Louis Anthony (Tony) Cox [a], Douglas Popken [a], M. Sue Marty [b], J. Craig Rowlands [b], Grace Patlewicz [c], Katy O. Goyak [d], Richard A. Becker [e,*]

[a] Cox Associates, 503 Franklin St, Denver, CO 80218, USA
[b] The Dow Chemical Company, Toxicology & Environmental Research & Consulting, 1803 Building Washington Street, Midland, MI 48674, USA
[c] DuPont Haskell Global Centers for Health and Environmental Sciences, Stine 320/212, 1090 Elkton Road, Newark, DE 19711, USA
[d] ExxonMobil Biomedical Sciences Inc., 1545 Route 22, East Annandale, NJ 08801, USA
[e] Regulatory and Technical Affairs Department, American Chemistry Council (ACC), 700, 2nd Street NE, Washington, DC 20002, USA

ARTICLE INFO

Article history:
Received 2 February 2014
Available online 17 May 2014

Keywords:
Validation framework
High throughput/high content assays
Prediction models
Adverse outcome pathways
Endocrine

ABSTRACT

High throughput (HTS) and high content (HCS) screening methods show great promise in changing how hazard and risk assessments are undertaken, but scientific confidence in such methods and associated prediction models needs to be established prior to regulatory use. Using a case study of HTS-derived models for predicting *in vivo* androgen (A), estrogen (E), thyroid (T) and steroidogenesis (S) endpoints in endocrine screening assays, we compare classification (fitting) models to cross validation (prediction) models. The more robust cross validation models (based on a set of endocrine ToxCast™ assays and guideline *in vivo* endocrine screening studies) have balanced accuracies from 79% to 85% for A and E, but only 23% to 50% for T and S. Thus, for E and A, HTS results appear promising for initial use in setting priorities for endocrine screening. However, continued research is needed to expand the domain of applicability and to develop more robust HTS/HCS-based prediction models prior to their use in other regulatory applications. Based on the lessons learned, we propose a framework for documenting scientific confidence in HTS assays and the prediction models derived therefrom. The documentation, transparency and the scientific rigor involved in addressing the elements in the proposed Scientific Confidence Framework could aid in discussions and decisions about the prediction accuracy needed for different applications.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

Advances in high throughput (HT) and high content (HC) screening methods for quantifying and characterizing molecular and cellular responses to chemicals are creating a new frontier in toxicity testing and subsequent hazard and risk assessments (Wetmore et al., 2012; Thomas et al., 2013). The National Academy report "Toxicity Testing in the 21st Century: A Vision and a Plan" (NRC, 2007) has been embraced by the U.S. Environmental Protection Agency (EPA) through its own computational toxicology research program. Over the last 6 years, the EPA has actively pursued its ToxCast™ program (http://www.epa.gov/ncct/toxcast/) and was instrumental in establishing the Tox21 collaboration of EPA, FDA, NIEHS/NTP and NIH (http://www.epa.gov/ncct/Tox21/), launched in 2008, "to develop ways to predict potential toxicity of chemicals and to develop a cost-effective approach for prioritizing the thousands of chemicals that need toxicity testing" (http://www.epa.gov/ncct/Tox21/).

Actualizing the Tox21 vision entails two main components: (1) HT/HC screening methods to construct quantitative biological profiles of chemical substances, and (2) prediction models to interpret these profiles in the appropriate context of hazard and risk assessment. Such prediction models are typically derived in one of three different approaches. One approach is to use HT/HC screening methods as inputs to predict *in vivo* animal toxicity study outcomes. A number of statistical ToxCast™ models have been developed to predict adverse outcomes in animal toxicity studies, including hepatocarcinogens (Shah et al., 2011), cancer (Kleinstreuer et al., 2012), reproductive toxicity (Martin et al., 2011) and prenatal developmental toxicity (Sipes et al., 2011). For the most part, these models have been based on correlating

* Corresponding author. Fax: +1 202 488 2503.
E-mail addresses: tcoxdenver@aol.com (L.A. (T.) Cox), doug@cox-associates.com (D. Popken), MSMarty@dow.com (M.S. Marty), JCRowlands@dow.com (J.C. Rowlands), patlewig@hotmail.com (G. Patlewicz), katy.o.goyak@exxonmobil.com (K.O. Goyak), Rick_Becker@americanchemistry.com (R.A. Becker).

http://dx.doi.org/10.1016/j.yrtph.2014.05.010
0273-2300/© 2014 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

or classifying HTS data with *in vivo* data extracted from ToxRefDB (http://www.epa.gov/ncct/toxrefdb/). Similarly, Fielden et al., 2011 developed an approach for using genomic signatures to predict nongenotoxic hepatocarcinogens in the rat.

A second alternative approach is to predict activity in specific biological pathways. The ToxCast™ program has published a number of models for endocrine disruption pathways (Reif et al., 2010), differentiation (Chandler et al., 2011) and angiogenesis (Kleinstreuer et al., 2013), and Andersen and colleagues have focused on pathways such as DNA damage and repair and PPARα (Bhattacharya et al., 2011; McMullen et al., 2014). A third approach could be termed a quantitative HT/HC screening-level chemical safety assessment. In this third approach, rather than predicting any specific hazard or pathway effect, HT/HC screening tools are coupled with exposure information/data to derive putative safe margins of exposure. This last approach appears to follow the vision articulated in the NAS 2007 publication most closely. Thomas et al. (2013) have used ToxCast™ assays and transcriptomics from short-term *in vivo* studies used to set benchmark doses without constraining the specific gene responses to those necessarily linked to known adverse outcomes. These benchmark doses are then compared to estimates of human exposures to derive margin of exposure metrics for determining safety.

Thus, understanding and evaluating the predictivity of HTS/HCS models is an important aspect of regulatory acceptance and use of Tox21 approaches. This can be both challenging and controversial. Evaluation of model performance can be characterized, in part, by evaluating sensitivity (positive predictivity) and specificity (negative predictivity). Balanced accuracy (BA) is the arithmetic mean of sensitivity and specificity; the closer a value is to unity, the greater the accuracy. Thomas et al. (2012) analyzed the suite of endpoints in ToxRefDB, and reported (1) most *in vivo* endpoints could not be predicted any better using the HT/HC ToxCast™ assay data compared with chemical descriptor information (e.g., typical inputs used in the development of QSAR models); (2) the ability of ToxCast™ assays to predict most endpoints resulted in balanced accuracies of less than 0.55; (3) past estimates of predictive performance of models developed from ToxCast™ assays were likely impacted by "prefiltering the *in vitro* assay data outside the cross-validation loop," in other words, manually selecting a subset of chemicals which were expected to be relatively predictable as the basis for assessing predictive power. Dix and colleagues responded that failing to incorporate knowledge of biological or adverse outcome pathways needlessly limits the purely statistical approach to classification, making its predictive power less than it would and should be if relevant biological, chemical, and toxicological knowledge were used to select relevant predictors (Dix et al., 2012). Clearly, greater attention needs to be drawn to how HTS/HCS derived prediction models are constructed and evaluated.

2. Case study: evaluation of estrogen, androgen and thyroid prediction models of Rotroff et al. (2013)

To better understand the impact of classification versus modeling using cross validation, we used the endocrine screening prediction models published by Rotroff et al. (2013) as a case study. This example was of particular interest given the EPA's plan to enhance the efficiency of its Endocrine Disruption Screening Program (EDSP) with HTS assays from the ToxCast program to achieve the vision described as EDSP21(EPA, 2011). EPA's current EDSP (http://www.epa.gov/endo/) is a two tiered screening and testing program. EDSP Tier I (EDSP T1S) consists of a battery of *in vitro* and *in vivo* screening assays designed to evaluate the potential of chemicals to interact with the Estrogen, Androgen or Thyroid (EAT) hormone systems. EDSP Tier II is comprised of *in vivo* tests which are intended to be used for Tier 1 positive substances to evaluate adverse effects, NOAELs and dose response for use in risk assessment (EPA, 2009). Under EDSP21, EPA plans to use HTS assays and prediction models first for priority setting, then to replace certain EDSP T1S assays and finally, if scientifically supportable, to fully replace EDSP T1S altogether (EPA, 2011).

To select a set of substances with robust data, Rotroff et al. (2013) used the data from two sources: (1) guideline study data generated during the validation of several EDSP *in vivo* endocrine screening assays (e.g., uterotrophic, Hershberger assay, intact male, fish short-term reproduction, and the male and female pubertal assays, etc.) and (2) non-guideline endocrine-related studies from the open literature. For the same substances, relevant *in vitro* ToxCast™ endocrine screening assay results were used as the basis for construction of a set of models to predict results for *in vivo* estrogen receptor-, and androgen receptor-mediated responses in the EDSP T1S *in vivo* assays (e.g., E responses in the uterotrophic, fish short-term reproduction and pubertal female assays and A responses in the Hershberger, fish short-term reproduction and pubertal male assays). Rotroff et al. (2013) reported that "ToxCast™ estrogen receptor and androgen receptor assays predicted the results of relevant EDSP T1S assays with balanced accuracies of 0.91 ($p < 0.001$) and 0.92 ($p < 0.001$), respectively," concluding that "[O]verall, results suggest that current ToxCast™ assays can accurately identify chemicals with potential to interact with the estrogenic and androgenic pathways, and could help prioritize chemicals for EDSP T1S assays."

This case study aimed to (1) replicate the prediction models derived by Rotroff et al. (2013) (referred to below as the "EPA 2013 HTS-derived EDSP T1S endocrine prediction models"); (2) evaluate the differences between model development using fitting and model development using cross validation; and (3) explore the feasibility of deriving prediction models for EDSP outcomes either using the full range of ToxCast™ assays or chemical specific information (e.g., classical QSAR). Based upon the knowledge gained, we describe a framework that may be useful to explicitly document scientific confidence in HTS/HCS assays and prediction models derived therefrom, and discuss the case study in terms of the elements of this framework.

2.1. Data and methods

In vitro assay data and *in vivo* outcome data were obtained from the Supplemental Data files provided in the online version of Rotroff et al. (2013): Supplemental File 1 contained the data from the ToxCast™ assays, while Supplemental File 3 contained the data from the EDSP assays. Although there were over 500 individual ToxCast™ assays in the EPA database, only 18 ToxCast™ assays of three distinct types (i.e., competitive binding, enzyme inhibition, and reporter gene assays) were included in Supplemental File 1, with specific subsets of these applied to each of the four EDSP assay classes: estrogen-, androgen-, steroidogenesis-, or thyroid-related MOAs (guideline-E, guideline-A, guideline-S, guideline-T, respectively). Although Rotroff et al. (2013) also compared the limited set of ToxCast™ thyroid-related results and the one ToxCast™ steroidogenesis assay with guideline and non-guideline studies, the underlying datasets were so sparse that constructing and evaluating prediction models had limited utility. Rotroff et al. (2013) thus concluded there is a "clear need to develop HTS assays capable of detecting steroidogenesis and thyroid disrupting compounds." Therefore, although we have included thyroid and steroidogenesis for completeness in our case study, we do not discuss the results in any great detail.

To predict EDSP T1S outcomes from ToxCast™ and/or chemical predictors, a Random Forest machine learning algorithm, available in the R statistical computing environment (http://cran.

r-project.org/web/packages/randomForest/index.html) was used. Random Forest is an advanced predictive algorithm that combines results from multiple classification trees (a statistical algorithm that selects the most predictive variables and ranges for the outcome of interest), (Brieman, 2001). The regression mode of Random Forest was used to directly fit to the FractionPositive (see Section 3.2 for definition) decimal values. *Model fitting* was performed akin to that undertaken in Rotroff et al. (2013) and *model prediction* (cross validation) was performed similar to Thomas et al. (2012). In the *model fitting* approach, all the data (guideline/non-guideline in turn) were used to both train the model (i.e., estimate regression coefficients to optimize measures of fit), and obtain metrics of the quality of the fit obtained. By using the same data sets to train the model and obtain metrics, model overfitting can artificially inflate the prediction values. Therefore, in the *model prediction* approach, a 5-fold cross-validation was carried out to reduce the overfitting bias. For cross validation, 80% of the data was selected at random to train the model, yielding a regression model for predicting EDSP outcomes from ToxCast™ predictors. The resulting model was then tested on the 20% of the data not used in developing the model. This was done for each 20% block of data in turn, and the entire process was repeated for each of 100 random permutations of the data set, generating a total of 500 different comparisons of model-based predictions to true EDSP outcomes. The performance metrics, such as the mean squared error for the difference between model-predicted and observed values of positive EDSP tests (i.e., *FractionPositive*, see definition below), were averaged over all 100 full prediction sets to obtain the final results. To obtain classification results and balanced accuracy (BA) values, final predicted values and the known *FractionPositive* values were dichotomized. (All datasets and R software files are available in Supplemental Material or may be obtained by contacting T. Cox directly.)

2.2. Re-creation and evaluation of the Rotroff et al. (2013) HTS-derived EDSP T1S endocrine prediction models

2.2.1. Data set construction

Recreating the Rotroff et al. (2013) HTS-derived EDSP T1S endocrine prediction models required reconstructing their datasets, including following how dose–response data from the ToxCast™ assays were summarized and classified (e.g., as positive or negative responses). From the results of the 18 specific ToxCast™ assays, Rotroff et al. (2013) assessed how well counts of positive tests could be used to classify, or fit, EDSP T1S *in vivo* results (i.e. fraction that were positive), compared with using count values from randomly selected subsets of the 18 ToxCast™ assays. Positive hit counts were their sole predictor, and thresholds for classifying the number of "positive" hit counts (i.e., one positive E-related assay, two positive E-related assays, etc.) were chosen to maximize balanced accuracy. Full details of how the dose–response data were processed were not given in the Supplementary Materials, and despite several helpful discussions with the authors of the HTS-derived EDSP T1S endocrine prediction models, we could not replicate them. We therefore initially accepted the chemical-assay binary "hit" values as provided, using them as inputs to data mining and machine learning software available in the R statistical computing environment.

To create our own data sets relating predictors to EDSP results, we merged the ToxCast™ chemical-assays and EDSP result files with a matching key (CAS registration number and Class). Comparing the resulting dataset against the supporting information/data provided in Rotroff et al. (2013) revealed one additional chemical, benomyl, as a candidate for inclusion in the guideline A set; i.e., benomyl appeared in both the original ToxCast™ and EDSP data files for A-related endpoints, and was not filtered out via any of the exclusion criteria stated in Rotroff et al. (2013). Our analysis therefore began with the same data used by Rotroff et al. (2013) as far as possible, with the exception of this additional chemical.

Table 1 shows an example of the resulting data for the first 3 variables for guideline-E combinations. Note: results for all chemical-class combinations, including hit indicators, for both guideline and non-guideline EDSP cases are provided in the Supplementary Files.

Although the preceding data preparation steps closely followed Rotroff et al. (2013)'s general approach, the following differences are reflected in predictors (*CountTC* and *FractionTC*) and predicted variables (*FractionPositive*) for guideline and non-guideline studies in our study: (1) we considered both *FractionTC* and *CountTC* as predictors, allowing a Class-independent input which was considered necessary since each *Class*, E, A, S, and T contained different numbers of chemical-assay pairs; (2) instead of dichotomizing *FractionPositive* into positive or negative classes, we provided the raw decimal values to the prediction algorithms. For example, if three out of the four E-related assays were positive, the value of 0.75 was used rather than 1, which had been taken to designate positive results in Rotroff et al. (2013); (3) we considered individual ToxCast™ assay results, as well as their collective counts. Each result was retained; and (4) all classes were modeled simultaneously to enable sharing of potentially predictive information. This also resulted in larger sample sizes. However, we performed separate analyses for guideline and non-guideline EDSP studies, as in Rotroff et al. (2013).

2.2.2. Evaluation of the EPA 2013 HTS-derived EDSP T1S endocrine prediction models

The challenge was therefore to determine how well *FractionPositive* for guideline and non-guideline studies could be predicted from *CountTC*, *FractionTC*, and the 18 assay specific hit indicators within each EDSP class grouping. Exploratory analysis showed that *CountTC* and *FractionTC* are significantly correlated with *FractionPositive* for the EDSP assays, (both guideline and non-guideline studies), even though most individual ToxCast™ assays are significantly correlated with non-guideline EDSP results. We applied a Random Forest machine-learning algorithm to the guideline and non-guideline data sets described above to classify or predict the EDSP *FractionPositive* values. The results are shown in Table 2. For model-fitting, the results are comparable to those of Rotroff et al. (2013). Table 2 includes the number wrong (#Wrong) [number of instances incorrectly classified using the fitted model], balanced accuracy (BA), the root mean squared error (RMSE) (which Rotroff et al., 2013 did not report) and also the corresponding model-prediction results based on model cross-validation (CV) using disjoint training and test sets, as previously described, with metrics from cross-validation (denoted by CV) averaged over 100 random CV prediction sets.

The fitted results for guideline studies (Table 2; using ToxCast™ indicators to classify known EDSP results) compare favorably to the fitted results obtained by Rotroff et al. (2013) (BA = 0.94 using Random Forest versus 0.91 obtained by Rotroff et al. for the E class, reflecting one additional correct classification by Random Forest; identical results to Rotroff et al. for the A and T classes; and slightly lower BA for the S class). As expected, the true *prediction* task, using cross validation, is more challenging. For E and A, the BA values for prediction (BACV) are lower than the BA values for fitting (BA) by approximately 0.09 and 0.13, respectively. (As noted by Rotroff et al. (2013), S and T responses could not be fitted very well due to limitations with these data sets (i.e., too few assays and/or insufficient positive results). The RMSE is also much higher for predicted values as compared to the fitted values. Thus, the very high BA values (>0.9) reported by Rotroff et al. (2013) for classes E and A are higher than the cross-validation results we derived. Consistent

Table 1
Example of initial predictor variables and dependent variable (guideline-E combinations shown).

CASRN	Name	Class	CountTC	FractionTC	FractionPositive
104-40-5	4-Nonylphenol	E	1	0.13	0.50
140-66-9	4-(Tert-octyl)phenol	E	4	0.67	1.00
521-18-6	5a-Anderostan-17B-ol-3-one	E	3	0.50	1.00
1912-24-9	Atrazine	E	0	0.00	0.33
17804-35-2	Benomyl	E	0	0.00	0.25
50-28-2	b-Estradiol	E	4	0.67	1.00
80-05-7	Bisphenol-A	E	6	0.75	0.90
2425-06-1	Captafol	E	0	0.00	0.33
5598-13-0	Chlorpyrifos-methyl	E	2	0.25	0.00
84-74-2	Dibutyl phthalate	E	1	0.13	0.00
117-81-7	Diethylhexyl phthalate	E	0	0.00	0.33
66230-04-4	Esfenvalerate	E	1	0.13	0.00
60168-66-0	Fearimol	E	3	0.38	1.00
13311-84-7	Flutamide	E	0	0.00	0.00
446-72-0	Genistein	E	4	0.67	1.00
72-43-5	Methoxychlor	E	3	0.38	1.00
40487-42-1	Pendimethalin	E	2	0.25	1.00
52645-53-1	Permethrin	E	1	0.13	0.00
67747-09-5	Prochloraz	E	1	0.13	1.00
3380-34-5	Triclosan	E	0	0.00	0.00
50471-44-8	Vinclozolin	E	1	0.13	0.00

Legend: CountTC is the total count (number) of the 18 ToxCast predictors that were classified by Rotroff et al. as providing a positive result, for each chemical and each EDSP assay class. For example, 7 would be entered if 7 of the 18 ToxCast assays were deemed positive for chemical X. *FractionTC* is the fraction of positive ToxCast results, calculated using only relevant (non-NA) assays in the denominator. Rotroff et al. (2013) selected subgroups of the 18 ToxCast assays as being relevant for predicting the results in each of the 4 EDSP assay classes (estrogen, androgen, steroidogenesis and thyroid (E, A, S, T)), such as 5 specific assays for androgen, and 8 or 9 for estrogen (depending upon the chemical). If, for example, 7 of 9 ToxCast estrogen-related relevant assays were positive for chemical X, 0.78 would be entered. *FractionPositive* is the fraction of all EDSP studies classified by Rotroff et al. (2013) as showing a positive result. This fraction was calculated separately for each of the 159 chemicals and within each of the 4 classes of EDSP assays (E, A, S, T) for which data were available. (For example, *FractionPositiveG* for chemical X is the proportion of E-related positive studies (the uterotrophic, fish short-term reproduction assay, ER binding, pubertal female, etc.) divided by the total number of all guideline-E studies in the data set, and *FractionPositiveNG* for chemical X is the proportion of E-related positive studies (the uterotrophic, fish short-term reproduction assay, ER binding, pubertal female, etc.) divided by the total number of all non-guideline-E studies in the data set.) Hit indicators were the 18 variables indicating the binary result for each chemical-assay combination. A "1" indicated a positive result, "0" a negative result, and "−1" indicated cases where a result was not relevant (NA in the original files).

Table 2
Fitting and prediction results for guideline and non-guideline studies using the 18 Selected ToxCast™ assays from the EPA 2013 HTS-derived EDSP T1S endocrine prediction models publication.

Class	Count	Model-fitting results based on original Rotroff data			Model-prediction results based on model cross-validation		
		# Wrong	BA	RMSE	# Wrong CV	BACV	RMSECV
Guideline studies							
E	21	1	0.9444	0.2092	2.98	0.8493	0.3431
A	14	1	0.9167	0.2159	2.72	0.7883	0.3352
S	17	6	0.5000	0.3902	6	0.500	0.4215
T	8	4	0.5000	0.5001	6.16	0.230	0.5717
Non-guideline studies							
E	143	12	0.7209	0.2584	16.05	0.6442	0.2946
A	59	19	0.5000	0.3931	21.38	0.4926	0.4149
S	55	25	0.5000	0.4157	25.33	0.4959	0.4301
T	47	16	0.5000	0.3853	16.38	0.4954	0.4046

Legend: Class = type of assay; Count = number of instances; # Wrong = number of instances incorrectly classified using the fitted model; BA = balanced accuracy metric for the fitted model; RMSE – root mean squared error between the actual FractionPositive and the model fit; # WrongCV = number of instances incorrectly classified by cross-validation prediction; BACV = balanced accuracy from cross-validation prediction; RMSECV = RMSE from cross-validation prediction. For BA and BACV, the closer a value is to unity, the greater the model accuracy. For RMSE and RMSECV, the smaller the value, the better predictive performance of the model.

with the concerns of Thomas et al. (2012), this suggests an upward bias in Rotroff et al.'s reported classification performance (BA) estimates compared to the results obtained using disjoint training and test results (BACV estimates). This outcome is a likely consequence of over-fitting (arising from the use of the same data to train and then test classification of known EDSP results using ToxCast™ results). Notwithstanding this finding, even when cross-validation is used, the resulting BA scores (i.e., BACV in Table 2) of 0.85 for Class E and 0.79 for Class A are significantly better than random and also well above the 0.55 level mentioned for most endpoints by Thomas et al. (2012). Thus, the major finding in the EPA 2013 HTS-derived EDSP T1S endocrine prediction model publication – that class A and E results for guideline studies can be predicted with substantially better than random accuracy using the 18 selected ToxCast™ assays results – is supported by this cross-validation analysis. Overall, the BA scores for cross validation are lower than for fitting: 0.85 compared 0.94 for class E, and 0.79 compared to 0.92 for class A. For steroidogenesis (S) and thyroid (T), cross validation BA values were <0.5, indicating poor predictivity of the ToxCast™ assays, consistent with the findings of Rotroff et al. (2013).

For the much larger set of non-guideline studies (Table 2), the performance of the automatically fitted Random Forest model for class E is similar to that of Rotroff et al.'s hand-tuned model (BA of 0.72 versus 0.74), while for class A, the BA is worse (0.50 versus 0.65). Prediction has lower BA than fitting for Class E, but is almost the same for other classes, and RMSE is only slightly lower for fitting than for prediction. Again, consistent with EPA 2013

HTS-derived EDSP T1S endocrine prediction models, for non-guideline studies, BA values indicated poor predictivity for steroidogenesis and thyroid. Compared to guideline studies, non-guideline studies can vary greatly in data quality and study reliability, and criteria for determining positive and negative responses are not standardized, so the results may be, in large part, a reflection of these features. Publication bias may also be an important factor to consider; positive studies are more likely to be published, which could skew the available data sets. A focus on quality and reliability of the input data is warranted when constructing datasets for use in building prediction models.

2.3. Evaluation of prediction modeling using the full ToxCast assay dataset

We also conducted an evaluation of how well the EDSP Tier 1 outcomes for E, A and T endpoints could be automatically predicted from the full set of ToxCast™ assays, rather than only from the subset of 18 endocrine-related ToxCast™ assays in Table 1. Here the concentration–response data for the ToxCast™ chemicals were merged with the Supplemental File 3 data of Rotroff et al. (2013), using CASRN as the common key. The concentration–response data were contained within the following files, downloaded from the ToxCast web site (http://epa.gov/ncct/toxcast/data.html): (1) ACEA_PrimaryData_ToxCast-PhaseI_320Chemicals_03Feb2010.txt; (2) Attagene_PrimaryData_ToxCast-PhaseI_320Chemicals-plus-controls_29Jan2010.txt; (3) Bioseek_PrimaryData_ToxCast-PhaseI_320Chemicals_03Feb2010.txt; (4) Cellumen_PrimaryData_ToxCast-PhaseI_320Chemicals-plus-controls_03Feb2010.txt; (5) CellzDirect_PrimaryData_ToxCast-PhaseI_320Chemicals_29Jan2010.txt; (6) Gentronix_PrimaryData_ToxCast-PhaseI_320Chemicals_14Dec2009.txt; and (7) NCGC_PrimaryData_ToxCast-PhaseI_320Chemicals_03Feb2010.txt.

For each matching chemical and each available ToxCast assay, a binary "hit" indicator (0 = no, 1 = yes) of positive activity, was independently generated. Briefly the procedure was to use the *drm*() (dose–response model) function of the DRC function library in the R programming environment (http://cran.r-project.org/web/packages/drc/) to approximate the raw concentration–response data for each chemical and ToxCast™ assay by a continuous curve. The resulting curves were passed to the *noEffect*() function of the same library to dichotomize them. The *noEffect*() function performs a significance test comparing the given dose–response curve to a simple linear regression with zero slope. If the function returned a p value less than .05, a "hit" was recorded. To make predictions with the full set of ToxCast™ assay data, these binary assay indicators were used instead of *CountTC*, and *FractionTC*.

The main results of applying the Random Forest algorithm to the full set of HTS (ToxCast™ assay) indicators are presented in Table 3. Significant improvements in fit are noted, with BA scores of 1 for classes of guideline-E and -A studies (Table 3), compared to 0.94 and 0.92 (Table 2) achieved using 18 endocrine assays for prediction model development. Significant improvements in BA for classes S and T were achieved as well. Thus, looking at BA, fully automated methods gave significantly better BA results than manual selection for both guideline and non-guideline studies, and for all classes. Theoretically, since the E and A responses in the uterotrophic and Hershberger assays are receptor mediated, one might have expected the opposite results. That is, by including all ToxCast™ assay results, the balanced accuracy for fitting would decrease, compared to models developed only with E- and A-specific ToxCast™ assays. However, since this is a classification model and ToxCast™ assays are being evaluated against known outcomes, there are likely elements within the features of the full suite of ToxCast™ assays that provide auxiliary value which enables improved classification.

Along such lines, it is critically important to note that despite their excellent performance, as measured by BA, in classifying known EDSP outcomes based on known ToxCast™ results, the fully automated methods do not achieve comparably high genuine predictive power, as indicated by their much lower BACV scores (Table 3). Thus, it is apparent that (1) state-of-the-art machine learning methods exhibit excellent ability to generate classification rules (fitting) that describe known data points (BA > 0.9 in Table 3, similar to Rotroff et al.'s findings), but only limited ability to predict the outcomes of EDSP assays which have not yet been performed (and/or which have not been included in the fitting exercise) (BACV < 0.6 for guideline studies and <0.7 for non-guideline studies); and (2) the BA scores for EPA 2013 HTS-derived EDSP T1S endocrine prediction models, while reflecting good fits to known data points, should likewise not be interpreted as necessarily indicating comparably high ability to actually predict EDSP results for tests which have not yet been performed.

2.4. Evaluation of prediction modeling using chemical property information

For completeness, an evaluation of how well EDSP outcomes could be predicted from chemical information alone was also undertaken. Chemical structure data available for the 159 ToxCast™ chemicals was extracted from the ToxCast™ website (http://epa.gov/ncct/toxcast/data.html). The file, ToxCast_Phase_1_StructureClassifiers_20110110.txt, contains binary structure indicators for 241 ToxCast™ chemical attributes. The structure data set was merged with the data files described above. To make predictions with structure data, these 241 binary structure indicators were used instead of *CountTC*, and *FractionTC* in Table 2. The Random Forest algorithm was applied to the chemical structure data as predictors. Table 4 shows the summary results.

The Random Forest algorithm provided an excellent *fit* using chemical structure data only (BA ranged from 0.70 to 1.0), especially for guideline studies. However, comparing Tables 3 and 4, the *predictions* based on chemical structures are essentially the same in quality as those based on using all ToxCast™ assay results (BACV ranged in Table 3 from 0.32 to 0.67 and in Table 4 from 0.31 to 0.67). Using the combination of all ToxCast™ predictors and all chemical structure predictors did not significantly improve the results over using either one alone (results not shown). These findings confirm the finding of Thomas et al. (2012) that use of HTS data does not greatly improve upon what can already be predicted from chemical property data alone.

3. Conclusions regarding the EPA 2013 ToxCast HTS-derived EDSP T1S endocrine prediction models

Our work to independently replicate and verify previously reported results (Rotroff et al. (2013)) that showed high balanced accuracy (BA > 0.90) values for model fitting to predict E and A results from selected ToxCast™ assay results has led to the following major conclusions:

(1) Although it was not possible to exactly replicate the model results of Rotroff et al. (2013) based on the documentation and data files provided, using the same 18 selected ToxCast™ assays as classifiers, BA values in excess of 90% for E and A could be achieved by model fitting of known EDSP results for 35 guideline studies (21 in Class E and 14 in Class A).
(2) Using fully automated machine-learning techniques that are publically available in the R statistical computing environment allows these model fitting BA scores to be improved to 100% for E and A for these same 35 chemicals.

Table 3
Fitting and prediction results for guideline and non-guideline studies using all ToxCast™ assays.

Class	Count	Model-fitting results based on original Rotroff data			Model-prediction results based on model cross-validation		
		# Wrong	BA	RMSE	# Wrong CV	BACV	RMSECV
Guideline studies							
E	15	0	1.0000	0.2651	6.18	0.5695	0.4607
A	14	0	1.0000	0.2551	5.72	0.5873	0.4765
S	17	2	0.8333	0.2303	6.9	0.5830	0.4740
T	8	1	0.8750	0.3828	5.45	0.3188	0.6366
Non-guideline studies							
E	135	3	0.9063	0.1265	16.08	0.5216	0.3007
A	59	2	0.9474	0.2343	16.31	0.6744	0.4253
S	55	7	0.2291	0.2291	19.39	0.6348	0.4216
T	47	8	0.2217	0.2217	17.91	0.5561	0.4154

Legend: Class = type of assay; Count = number of instances; # Wrong = number of instances incorrectly classified using the fitted model; BA – Balanced accuracy metric for the fitted model; RMSE – root mean squared error between the actual FractionPositive and the model fit; # WrongCV = number of instances incorrectly classified by cross-validation prediction; BACV = Balanced accuracy from cross-validation prediction; RMSECV = RMSE from cross-validation prediction. For BA and BACV, the closer a value is to unity, the greater the accuracy. For RMSE and RMSECV, the smaller the value, the better prediictive performance of the model.

Table 4
Fitting and prediction results for guideline and non-guideline studies using chemical structure data.

Class	Count	Model-fitting results based on original Rotroff data			Model-prediction results based on model cross-validation		
		# Wrong	BA	RMSE	# Wrong CV	BACV	RMSECV
Guideline studies							
E	15	0	1.0000	0.2377	4.89	0.5960	0.4400
A	14	0	1.0000	0.2318	4.24	0.6708	0.4577
S	17	3	0.7500	0.2222	6.58	0.5058	0.4671
T	8	0	1.0000	0.3772	5.53	0.3088	0.6000
Non-guideline studies							
E	135	6	0.8395	0.1674	16.78	0.6144	0.2961
A	59	4	0.8947	0.2579	16.85	0.6012	0.4488
S	55	11	0.7867	0.2773	26.26	0.5048	0.4616
T	47	10	0.7026	0.2392	19.15	0.4716	0.4182

Legend: Class = type of assay; Count = number of instances; # Wrong = number of instances incorrectly classified using the fitted model; BA – Balanced accuracy metric for the fitted model; RMSE – root mean squared error between the actual FractionPositive and the model fit; # WrongCV = number of instances incorrectly classified by cross-validation prediction; BACV = Balanced accuracy from cross-validation prediction; RMSECV = RMSE from cross-validation prediction. For BA and BACV, the closer a value is to unity, the greater the accuracy. For RMSE and RMSECV, the smaller the value, the better predictive performance of the model.

(3) Assertion that EDSP outcomes for classes E and A can be forecast from ToxCast™ data with BA > 0.9 should be understood to mean that these outcomes can be fitted or classified with BA > 0.9, once they are known. However, the important practical challenge of true prediction of as-yet unknown EDSP outcomes from ToxCast™ data appears to be much more difficult.
(4) Using cross validation methods, the E and A prediction models developed using guideline studies yielded BACV results of 0.85 (E) and 0.79 (A), which are significantly better than random.
(5) There is a caveat: the activities evaluated for E and A are based on estrogen receptor and androgen receptor interactions; other modes-of-action that affect E and A signaling (e.g., hypothalamic–pituitary–gonadal axis effects) were not evaluated as part of this analysis or that of Rotroff et al. (2013).
(6) Overall, most likely due to limitations in the available assays and data sets, for S and T, using the 18 ToxCast™ assays, both classification and prediction were no better than random. Similar conclusions were reached by Rotroff et al. (2013).
(7) Somewhat surprisingly, for both guideline and non-guideline studies, using all of the ToxCast™ assays improved classification (BAs ranging from 0.77 to 1.0). However, this was not the case for predictivity using cross validation, where BA values hovered around random chance (BACVs ranged from 0.32 to 0.67). Substantially similar results were obtained when the Random Forest algorithm was applied to chemical structure data. These findings highlight the importance of understanding model development based on classification or fitting, versus model development using cross validation.

4. Discussion: developing scientific confidence in HTS/HCS assays and prediction models based on such assays

Significant research and development activities have focused on generating data from Tox21 methods and, to a lesser extent, prediction models derived from these assays. Yet, only limited attention has been paid to systematic approaches for developing and documenting scientific confidence in these assays and prediction models. Whether for priority setting, integrated testing strategies, product stewardship, or regulatory decision making, practical application of HTS/HCS assays and derived prediction models requires development of scientific confidence in the assays and the use of assay results. This is essential so that regulatory agencies, the regulated community, and the public can be assured that the use of this new kind of knowledge is both scientifically credible and relevant for decision making. The level of scientific confidence necessary for HTS assays and HTS-derived prediction models will certainly vary by use, such that, for example, a higher level of confidence would be required for a regulatory hazard

Table 5
Major components of a scientific confidence framework for prediction models derived from HTS assays.

Analytical validation	Assessment of the biological basis and analytical performance of assays. Each HTS/HCS assay should map to a defined mechanistic endpoint (e.g., the intermediate or key event in the mode of action or AOP). A defined chemical domain of applicability, and documentation of assay performance characteristics (reliability, sensitivity, and specificity) and transparent data sets (to enable independent verification) should be readily available.
Qualification	Assessment of the prediction model derived from the HTS/HCS screening assays. A defined algorithm for each prediction model is needed to ensure transparency. Appropriate measures of goodness-of-fit, robustness and predictivity of the prediction model need to be presented. Some prediction models may be quantitative, others may be qualitative. Known limitations of each prediction model should also be summarized. Prediction models should be characterized in sufficient detail to facilitate review, reconstruction and independent verification of results.
Utilization	Contextual and weight-of-evidence analysis of the use (qualitative or quantitative) of the prediction model for a specific purpose. This includes summarizing results of Analytical Validation and Qualification steps, defining the intended purpose of the prediction model and documenting/justifying applications, based on weight of evidence, where there is sufficient scientific confidence to support the use of the prediction model. The types of uses to which the prediction models could be applied include, but are not limited to: (1) priority setting, where the model is used to identify priority substances that will go on to more detailed evaluation; (2) screening level assessment of a biomarker, where the model is used as a surrogate data point for a biochemical endpoint or a biomarker; (3) integrated testing strategy, where the model is used to describe/predict a hazard property in lieu of conducting a traditional animal toxicity study or (4) to predict an adverse outcome.

assessment decision compared to a priority setting decision to select substances for more in-depth toxicity testing.

Previously, in Patlewicz et al. (2013), we discussed scientific confidence in the context of two validation frameworks, the IOM Framework (IOM, 2010) and the OECD Validation principles (2004), and adapted these for consideration in HTS/HCS assays. Similarly, Hartung et al. (2013) stated that "validating the mechanism of a (group of) toxicant(s) is the basis for mechanistic validation of tests that identify those toxicants," and proposed that the performance of Tox21 test methods could be addressed through "mechanistic validation" which focuses on biological pathways and is carried out in six steps: (1) articulation of the pathway ("the biological/mechanistic circuitry ("in the absence of xenobiotic challenge") leading to the hazard; (2) documenting the evidence, based on results in validated models, that reference chemicals that cause the hazard "perturb the biology in question;" (3) development of test(s) that reflect this biology; (4) verification that toxicants acting by this mechanism also do so in the test(s); (5) verification that antagonism or interference of the mechanism blocks or hinders positive test results. Although not a one-to-one match, this approach is, in essence, very similar to a well-informed adverse outcome pathway (AOP) that includes support for a causal link between the key event and adverse outcome (Ankley et al., 2010). While Judson et al. (2013) agree that increasing the use of well characterized reference compounds is important; they de-emphasize the need for cross-laboratory testing, stressing instead the importance of data availability to enable a "web-based, transparent, and expedited peer review process."

Based on Patlewicz et al. (2013), the analysis reported here and our recent efforts in developing AOPs, we propose a Scientific Confidence Framework for prediction models derived from HTS assays (Table 5). This proposed framework contains three inter-related elements: Analytical Validation, Qualification and Utilization. Together, these three elements should facilitate systematic, transparent, and objective evaluation and documentation of the scientific confidence of HTS/HCS assays and their prediction models. Addressing these elements may be challenging, although it is envisioned that the manner and extent to which these need to be addressed will be context dependent.

To illustrate application of the proposed Scientific Confidence Framework, we considered the Rotroff et al. (2013) report and the case study presented above. With respect to the first framework element, Analytical Validation, each of the 18 endocrine ToxCast™ assays has a defined mechanistic endpoint. In terms of accessibility to ToxCast™ data to enable independent verification, or refinement of existing prediction models or development of new models, EPA's ToxCast™ program is to be commended. All of the data were accessible from the ToxCast™ web site or from the Supplemental Materials in the Rotroff et al. (2013) publication. However, we did not review or analyze the chemical domains of applicability or documentation of assay performance characteristics (reliability, sensitivity, and specificity). Such a review is important, and requires a detailed and separate analysis outside the scope of this manuscript.

In terms of the second framework element, Qualification (assessment of the prediction model derived from the HTS/HCS screening assays), the Rotroff et al. (2013) HTS-derived EDSP T1S endocrine prediction models fell well short. None of the published prediction model algorithms were transparent, nor were the algorithms clearly defined and documented. Input from the first author was helpful, but significant efforts were still needed to reverse engineer the prediction models. For the most part, the underlying data sets were available to support this reverse engineering activity, however, the precise details of how dose–response data were re-coded as binary and response indicators (using an unspecified Hill function with some manual adjustment) were not documented or readily available in the Supplementary Materials. Thus, considerable unnecessary effort was required to enable independent verification. In terms of documentation of measures of goodness-of-fit, robustness and predictivity of the prediction models, Rotroff et al. (2013), provided estimates for predictivity of classification (fitting); but unbiased performance determinations (cross validation) for prediction of as-yet unknown EDSP outcomes from ToxCast™ data were not provided. Our analysis suggests that focusing the assay set to those of known biological relevance may increase the predictive performance (although not necessarily the fitting performance). With respect to use of datasets in model development, the difference between results of guideline studies vs. non-guideline studies should be noted. Based on our experience, we would recommend that toxicity study databases assembled for use in constructing predictive modeling should be curated by experienced toxicologists, with a particular focus on defining study quality and data reliability using a predetermined set of criteria that allow for identification of positive versus negative responses based on accepted toxicological principles. However, the models still must be used with caution because the domain of applicability – the structural space of substances used to construct these classification and prediction models – is currently limited to 21 chemicals for estrogen pathways and 14 for androgen pathways. Thus, research and development needs to continue to develop data sets and prediction models, before prediction models derived from HTS/HCS results can be used as standalone algorithms to predict *in vivo* endocrine endpoints.

Regarding the third element of the framework, Utilization, Rotroff et al. (2013) indicate the intent is to use these models to "predict components of EDSP T1S and related results." First, we agree with Rotroff et al. (2013), that current ToxCast results are inadequate to predict S and T responses. For E and A, our analysis shows that because these models were developed using fitting, they fall short in truly predicting EDSP Tier 1 E and A *in vivo* results.

Nevertheless, the balanced accuracies of both the classification (fitting) and cross validation models for E and A activities signal significant advancement in developing biologically-based HTS/HCS-derived models for E and A endocrine activities. Yet, with respect to regulatory use, a number of questions still must be addressed concerning the acceptable range of predictivity (balanced accuracy) of an E or A prediction model necessary to use for a given purpose. How accurate must predictivity be for prioritizing substances for E and A for proceeding into EDSP screening? For bypassing certain E or A EDSP receptor-mediated *in vitro* Tier 1 assays? For use in predicting results of the uterotrophic and Hershberger assays, obviating the need for these to be carried out in Tier 1 of the EDSP? Is there a set degree of scientific confidence in the performance of the assays and prediction models, irrespective of the type of regulatory use, or, as we suggested previously (Patlewicz et al., 2013), is it accepted that different uses can allow different levels of uncertainty/confidence, and, if so, what are these, in quantitative scientific terms? How can integration of exposure data/estimates with such HTS activity-based measures/predictions (Wetmore et al., 2012; Becker et al., 2014) provide an improved context for priority setting and decision making? While such questions cannot be readily answered today, they warrant timely robust discussion within the regulatory science community. Clearly, with the caveat that the results to date reflect a limited number of substances, the ToxCast™-derived E and A classification and cross validation models may have considerable utility to discriminate between substances that are a priority for proceeding into EDSP Tier 1 screening and those that are not. For use in EDSP, as well as for other potential applications, the Scientific Confidence Framework we have proposed, or a similar tool, can assist in the transparent characterization and communication of the performance of both assays and prediction models.

Conflict of interest

The authors had complete control over the design, conduct, interpretation, and reporting of the analyses included in this manuscript. The contents of this manuscript are solely the responsibility of the authors and do not necessarily reflect the views or policies of their employers. Tony Cox and Douglas Popken received funding to support this research from the American Chemistry Council (ACC). Sue Marty, Craig Rowlands, Grace Patlewicz and Katy Goyak are employed by companies engaged in chemical product manufacturing. Richard Becker is employed by ACC, a trade association of U. S. chemical manufacturers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yrtph.2014.05.010.

References

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29, 730–741.

Becker, R.A., Hays, S.M., Kirman, C.R., Aylward, L.L., Wise, K., 2014. Interpreting estrogen screening assays in the context of potency and human exposure relative to natural exposures to phytoestrogens. Birth Defects Res. B Dev. Reprod. Toxicol. 101, 114–124.

Bhattacharya, S., Zhang, Q., Carmichael, P.L., Boekelheide, K., Andersen, M.E., 2011. Toxicity testing in the 21st century: defining new risk assessment approaches based on perturbation of intracellular toxicity pathways. PLoS One 6 (6), e20887.

Brieman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chandler, K.J., Barrier, M., Jeffay, S.C., Nichols, H.P., Kleinstreuer, N., Singh, A., Reif, D., Sipes, N., Judson, R., Dix, D.J., Kavlock, R.J., Hunter, E.S., Knudsen, T., 2011. Evaluation of 309 environmental chemicals using a mouse embryonic stem cell adherent cell differentiation and cytotoxicity assay. PLoS One 6 (6), e1840.

Dix, D.J., Houck, K.A., Judson, R., Kleinstreuer, N., Knudsen, T.B., Martin, M.T., Reif, D.M., Richard, A.M., Shah, I.A., Sipes, N., Kavlock, R.J., 2012. Incorporating biological, chemical and toxicological knowledge into predictive models of toxicity: Letter to the Editor. Toxicol. Sci. 130, 440–441.

EPA, 2009. Endocrine Disruptor Screening Program; Policies and Procedures for Initial Screening. Federal Register, vol. 74, No. 71, pp. 17560–17579. <http://www.epa.gov/endo/pubs/revised_pandp_frn_041509.pdf> (Wednesday 15.04.09)

EPA, 2011. Endocrine Disruptor Screening Program, Weight-of-Evidence: Evaluating Results of EDSP Tier 1 Screening to Identify the Need for Tier 2 Testing. <http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2010-0877-0021>.

EPA, 2013a. Prioritization of EDSP Universe of Chemicals Using Computational Toxicology Tools. <http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPP-2012-0818-0022>.

EPA, 2013b. Endocrine Disruptor Screening Program (EDSP), Weight-of-Evidence Process: Case Studies. <http://www.regulations.gov/#!docketBrowser;rpp=25;po=0;dct=SR;D=EPA-HQ-OPP-2013-0230>.

Fielden, M.R., Adai, A., Dunn II, R.T., Olaharski, A., Searfoss, G., Sina, J., Aubrecht, J., Boitier, E., Nioi, P., Auerbach, S., Jacobson-Kram, D., Raghavan, N., Yang, Y., Kincaid, A., Sherlock, J., Chen, S.-J., Car, B., on behalf of the Predictive Safety Testing Consortium, Carcinogenicity Working Group, 2011. Development and evaluation of a genomic signature for the prediction and mechanistic assessment of nongenotoxic hepatocarcinogens in the rat. Toxicol. Sci. 124, 54–74.

Hartung, T., Hoffmann, S., Stephens, M., 2013. Mechanistic validation. ALTEX 30, 119–130.

Institute of Medicine (IOM), 2010. Evaluation of biomarkers and surrogate endpoints in chronic disease. ISBN: 978-0-309-15129-0.

Judson, R., Kavlock, R.J., Martin, M.T., Reif, D., Houck, K.A., Knudsen, T.B., Richard, A.M., Tice, R.R., Whelan, M., Xia, M., Huang, R., Austin, C.P., Daston, G.P., Hartung, T., Fowle III, J., Wooge, W., Tong, W., Dix, D.J., 2013. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. ALTEX 30, 51–66.

Kleinstreuer, N., Dix, D.J., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Paul, K.B., Reif, D., Crofton, K.M., Hamilton, K., Hunter, R., Shah, I.A., Judson, R., 2012. In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. Toxicol. Sci. 131, 40–55.

Kleinstreuer, N.D., Dix, D.J., Rountree, M., Baker, N., Sipes, N., Reif, D., Spencer, R., Knudsen, T.B., 2013. A computational model predicting disruption of blood vessel development. PLoS Comput. Biol. 9 (4), e1002996.

Martin, M.T., Knudsen, T.B., Reif, D., Houck, K.A., Judson, R., Kavlock, R.J., Dix, D.J., 2011. Predictive model of rat reproductive toxicity from ToxCast high throughput screening. Biol. Reprod. 8, 327–339.

McMullen, P.D., Bhattacharya, S., Woods, C.G., Sun, B., Yarborough, K., Ross, S.M., Miller, M.E., McBride, M., Lecluyse, E.L., Clewell, R.A., Andersen, M.E., 2014. A map of the PPARα transcription regulatory network for primary human hepatocytes. Chem. Biol. Interact. 209, 14–24.

National Research Council (NRC), 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. National Academy Press, Washington, DC.

Patlewicz, G., Phillips, R.D., Rowlands, J.C., Seidel, S.D., Simon, T., Goyak, K., Becker, R.A., 2013. Use and validation of HT/HC assays to support 21st century toxicity evaluations. Regul. Toxicol. Pharmacol. 65, 259–268.

Reif, D., Martin, M.T., Tan, S., Houck, K.A., Judson, R., Richard, A.M., Knudsen, T.B., Dix, D.J., Kavlock, R.J., 2010. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environ. Health Perspect. 118, 2–8.

Rotroff, D.M., Dix, D.J., Houck, K.A., Knudsen, T.B., Martin, M.T., McLaurin, K.W., Reif, D.M., Crofton, K.M., Singh, A.V., Xia, M., Huang, R., Judson, R.S., 2013. Using in vitro high throughput screening assays to identify potential endocrine disrupting chemicals. Environ. Health Perspect. 121, 7–14.

Shah, I.A., Houck, K.A., Judson, R., Kavlock, R.J., Martin, M.T., Reif, D., Wambaugh, J.F., Dix, D.J., 2011. Using nuclear receptor activity to stratify hepatocarcinogens. PLoS One 6 (2), e14584.

Sipes, N.S., Martin, M.T., Reif, D.M., Kleinstreuer, N.C., Judson, R.S., Singh, A.V., Chandler, K.J., Dix, D.J., Kavlock, R.J., Knudsen, T.B., 2011. Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. Toxicol. Sci. 124, 109–127.

Thomas, R.S., Black, M.B., Li, L., Healy, E., Chu, T.M., Bao, W., Andersen, M.E., Wolfinger, R.D., 2012. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. Toxicol. Sci. 128, 398–417.

Thomas, R.S., Philbert, M.A., Auerbach, S.S., Wetmore, B.A., Devito, M.J., Cote, I., Rowlands, J.C., Whelan, M.P., Hays, S.M., Andersen, M.E., Meek, M.E., Reiter, L.W., Lambert, J.C., Clewell 3rd, H.J., Stephens, M.L., Zhao, Q.J., Wesselkamper, S.C., Flowers, L., Carney, E.W., Pastoor, T.P., Petersen, D.D., Yauk, C.L., Nong, A., 2013. Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. Toxicol. Sci. 136, 4–18.

Wetmore, B.A., Wambaugh, J.F., Ferguson, S.S., Sochaski, M.A., Rotroff, D.M., Freeman, K., Clewell 3rd, H.J., Dix, D.J., Andersen, M.E., Houck, K.A., Allen, B., Judson, R.S., Singh, R., Kavlock, R.J., Richard, A.M., Thomas, R.S., 2012. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicol. Sci. 125, 157–174.