Receiver Operating Characteristic Analysis for Environmental Diagnosis. A Potential Application to Endocrine Disruptor Screening: In Vitro Estrogenicity Bioassays

ARTICLE *in* ENVIRONMENTAL SCIENCE AND TECHNOLOGY · AUGUST 2005

Impact Factor: 5.33 · DOI: 10.1021/es048598o · Source: PubMed

Environ. Sci. Technol. **2005**, *39,* 5349−5355

Receiver Operating Characteristic Analysis for Environmental Diagnosis. A Potential Application to Endocrine Disruptor Screening: In Vitro Estrogenicity Bioassays

OLWENN V. MARTIN,† KA MAN LAI,‡ MARK D. SCRIMSHAW,† AND JOHN N. LESTER*,†

Environmental Processes and Water Technology Group, Department of Environmental Science, Imperial College London, London SW7 2AZ, and Department of Civil and Environmental Engineering, University College London, London, WC1E 6BT, U.K.

The realization that certain chemicals are able to disrupt hormonal systems in humans and wildlife has challenged the way we assess risk from chemicals and led national and international agencies to devise programs to screen chemicals for endocrine-disrupting properties. Chemicals capable of mimicking sex hormones, such as estrogens and androgens, have received the most attention, and although not yet validated, in vitro techniques to test for such properties are well developed. Receiver operating characteristic (ROC) analysis has been successfully used in the biomedical and military fields for several decades to assess the accuracy of diagnostic tests in terms of both their sensitivity and specificity. This approach is applied here to demonstrate its potential to assess how well in vitro bioassays can predict estrogenicity in vivo. Despite the limited availability of suitable data, the ROC curves obtained indicate that these bioassays are effective diagnostic tests. The potential sources of false positives and false negatives are identified and potential applications to endocrine disruptor screening programs discussed.

Introduction

The ability of natural and man-made chemicals to mimic hormones and/or modulate hormonal responses is currently one of the most controversial environmental issues. The endocrine disruption hypothesis emerged from parallel observations in humans and wildlife; the higher incidence of certain cancers of the reproductive tract and other reproductive or developmental effects were linked to the prescription of the synthetic estrogenic drug diethylstilbestrol during pregnancy and in utero exposure, while reproductive and developmental effects in the progeny of wildlife exposed to higher concentrations of chemicals were recorded (*1*). Although the concept of endocrine disruption first developed when it was observed that some environmental chemicals were able to mimic the action of the sex hormones estrogens and androgens, it has now evolved to encompass a range of mechanisms involving the many hormones secreted directly into the blood circulatory system by ductless glands and their specific receptors distributed nonuniformly in diverse tissues (*2*). While the androgenic effects of tributyltin on mollusks have been well documented and resulted in population level effects, there is still debate over the ecological significance of the estrogenic effects of sewage treatment works discharges on fish observed more recently (*3*, *4*). While it is argued that the relative potencies of xenoestrogens are very often several orders of magnitude lower than those of endogenous hormones, there is emerging evidence that they can act additively or synergistically and that even low concentrations of weakly estrogenic compounds could contribute to the overall effect (*5*).

These concerns have led many national and international agencies to devise large programs to screen chemicals for such potential effects. The ultimate goal of these screening programs is a hazard assessment to help evaluate the risk posed by the presence of these chemicals in the environment. The European Union and the USEPA have both launched their priority lists of chemicals based on different criteria (*6*, *7*), while validation programs for the range of in vitro bioassays developed to detect endocrine-disrupting potential are ongoing (*8*, *9*). As such assays measure effects at the molecular and/or cellular levels, they are specific to one mode of action and essential to elucidate the mechanistic basis of body level effects (*10*). Estrogenicity is a term encompassing a range of effects observed in vivo and is believed to be the result of a number of possible modes of action, one of which is mediated by the estrogen receptor (ER). As a result, several in vitro bioassay techniques to test for estrogen agonists (or antagonists), such as the ER ligand binding assay, estrogen responsive element (ERE) reporter gene transcriptional assays including the yeast estrogenicity screen (YES), and the cell proliferation assay (e-screen), have been developed. In vitro bioassays are rapid, cost-effective tools that can achieve lower detection limits than chemical analysis (*11*). However, it is clear from the definition adopted during a major European workshop ["an endocrine disrupter is an exogenous substance that causes adverse health effects in an intact organism, or its progeny, subsequent to changes in endocrine function" (*12*)] that endocrine-disrupting effects have to be verified at the body level. These should be distinguished from potential endocrine disruptors shown to be active in vitro ["a *potential* endocrine disrupter is a substance that possesses properties that might be expected to lead to endocrine disruption in an intact organism" (*12*)]. Tens of thousands of chemicals need to be screened for potential endocrine disruption, although how effective these large screening programs will prove at detecting hazardous chemicals will be influenced by the diagnostic quality of the in vitro bioassays in predicting effects in vivo. Understanding how these assays perform should help tailor screening sequences that can avoid both misdiagnosis and duplication of effort.

Receiver operating characteristic (ROC) analysis has been commonly used in the biomedical and military fields to assess the performance of a diagnostic test for many years. It was recently proposed by Shine and co-workers (*13*) as a method to evaluate sediment quality guidelines for metals. It allows the analysis of the accuracy of a diagnostic test in terms of both its sensitivity (probability to correctly identify positive cases) and its specificity (probability to correctly exclude negative cases). Accuracy as a performance indicator represents the ratio of correct decisions achieved by a diagnostic test but does not reveal the type of incorrect decisions. The consequences of a false negative diagnosis and those of a

* Corresponding author phone: +44 (0)20 7594 6014; fax: +44 (0)20 7594 6016; e-mail: j.lester@imperial.ac.uk.
† Imperial College London.
‡ University College London.

10.1021/es048598o CCC: $30.25
Published on Web 06/09/2005

false positive diagnosis could be quite different in given circumstances, and accuracy is of limited usefulness. ROC analysis is able to decipher diagnostic performance in terms of both types of error and therefore better inform decisions taken on the basis of the result from a given test (*14*). This technique is adopted in this paper to illustrate its potential application in environmental diagnosis as a means of measuring the validity of decisions taken regarding possible effects of chemicals on the basis of data from in vitro assays. In vitro estrogenicity bioassay results from the literature will be assessed with respect to their ability to induce estrogenicity in vivo, on the basis of the results of the rodent uterotrophic assay, also extracted from the literature.

Methodology

Estrogenicity was first defined as the ability of a compound to induce a physiological response in vivo known as estrus, characterized at the cellular level by cell proliferation and the hypertrophy of female secondary sex organs at the body level, as well as the synthesis and secretion of cell-type-specific proteins (*15*, *16*). It is understood as the result of a sequence of actions initiated by the binding of an estrogenic compound to the estrogen receptor causing the dissociation of a heat shock protein. The liganded receptor can then form a homodimer complex able to bind to specific DNA sequences responsive to estrogens and initiate gene transcription. Several in vitro bioassays have been developed with end points measured at different stages of this sequence of actions. However, this is a simplistic model, and in reality, two subtypes of the estrogen receptor, ERα and ERβ, have been identified, they have shown differing ligand binding and DNA binding domains, and they can lead to antagonistic biological effects (*17*). ERs are also found on cell membranes where they act via nontranscriptional pathways. Additionally, estrogenic effects can be induced via non-ER-mediated mechanisms involving plasma transport proteins and/or the disruption of the synthesis of the receptors, or their related enzymes.

To be able to undertake ROC analysis and determine the sensitivity and specificity of those assays, it is necessary to have a measure of estrogenicity that is independent of the assay used and that will determine whether any chemical is a "true" estrogen. In this work, data from the rodent uterotrophic assay was used as a true measure of estrogenicity. Therefore, for clarity purposes, compounds active in this assay will be referred to as estrogens and those only active in vitro as potential estrogens. The rodent uterotrophic assay is an in vivo gravimetric assay based on the hypertrophy of the uterus that therefore incorporates all aspects of the endocrine system, allowing for absorption, metabolism, distribution, excretion, and alternative pathways and is considered the "gold standard" for estrogenicity (*18*, *19*). The Organization for Economic Co-operation and Development (OECD) recently completed an international validation program of the rodent uterotrophic bioassay and found it to be predictive in practice and reproducible across laboratories (*20*). However, while validation should address discrepancies in practice such as route of exposure, age of exposure, diet, and rodent species and strains, other issues such as the intrinsic variability of tissue weight and reduced strain sensitivity over time may prove harder to address (*21*).

Dataset. Estrogenicity in vitro assays are not yet validated, and data from the rodent uterotrophic assay reported in the literature are still based on differing protocols or practices. Results expressed as relative potencies or ratios are expected to be less sensitive to such variations than results expressed as a concentration (*22*). The data collected were therefore limited to sources reporting a potency or affinity ratio relative to 17β-estradiol (E2) for compounds tested with the rodent uterotrophic bioassay and the ER ligand binding assay, YES, the human reporter gene assay, or the e-screen. Both the requirements to find a matching set of data for the uterotrophic assay and at least one in vitro assay and the diversity of ways in which data have been reported limited the size of the dataset available for the analysis (Table 1).

Construction of ROC Curves. Most diagnostic tests yield a single number as their result. The distributions of result values of both the actually positive and actually negative cases will generally overlap (otherwise the test is perfect and can discriminate between positive and negative cases without error). For policy purposes, a threshold value needs to be chosen to differentiate between positive and negative results. Different threshold values will lead to different sensitivity and specificity ratios. An ROC curve can be traced by plotting the sensitivity against 1 − specificity by varying this threshold level (*14*).

$$\text{sensitivity} = P(T^+|E^+)$$

$$\text{specificity} = P(T^-|E^-)$$

where $P(X)$ is the probability of an event X, T^+ is the number of results over the given threshold for chemicals having shown estrogenicity in vivo, T^- is the number of results over the given threshold for chemicals having shown no estrogenicity in vivo, E^+ is the number of chemicals in the dataset for each in vitro test that has shown estrogenicity in vivo, and E^- is the number of chemicals in the dataset for each in vivo test that has shown no estrogenicity in vivo.

Diagnostic truth must therefore be known, and the chosen criterion for actual evidence of estrogenicity in vivo is the rodent uterotrophic assay. ROC analysis can be used where a dichotomy can be defined. In this case, individual compounds need to be classified either as estrogenic or nonestrogenic. Any substance found to be active in the rodent uterotrophic assay was classified as estrogenic regardless of relative uterotrophic potency.

Both extreme thresholds are those for which all cases are considered to give a positive result in the test, the point with (1, 1) coordinates, and all cases are considered to give a negative result, the point with (0, 0) coordinates. The diagonal between these two points represents a useless test (cannot positively discriminate between positive and negative cases). Similarly, the closer the curve is to the upper left-hand corner of the graph (approaching sensitivity = 1 and specificity = 1), the better the test with optimal sensitivity and specificity (Figure 1).

Four or five threshold values between these extremes were chosen to trace the curves. They were set depending on the orders of magnitude found in the test results. Including more threshold levels would not improve the curves as each dataset included just over 10 chemicals.

The points obtained are experimental estimates of operating points on a single ROC curve, and the area under the curve is a representation of the diagnostic effectiveness of the test.

Results

ROC Analysis. *ER Binding Assay*. This assay is used to determine the ability of a given compound to compete with radiolabeled E2 for binding to the ER. It is based on the assumption that binding to the ER would result in a subsequent effect on biological activity. The dataset included phytosterols, styrene oligomers, natural estrogens, and analogues. Of the 14 chemicals, only 4 were estrogens, i.e., active in the rodent uterotrophic assay (Table 1, second and fourth columns). While this positive case incidence seems reasonably representative of a battery of chemicals to be screened for estrogenicity, the dataset is not really diverse

TABLE 1. In Vivo and in Vitro Assay Data[a]

	rodent uterotrophic assay		ER binding assay		YES		human reporter gene assay		e-screen	
compd	RP (%)		RBA (%)		RP (%)		ln(PC10)		RPE (%)	
17β-estradiol (E_2)	100	(*23*)	100	(*24*)	100	(*23*)	5.36	(*25*)	100	(*26*)
17α-estradiol	50	(*27*)	50	(*27*)	5.25	(*23*)				
17-desoxy-E_2	12.5	(*27*)	56	(*27*)	10	(*27*)				
bidesoxy-E_2	10	(*27*)	0.1	(*27*)	0.01	(*27*)				
6-hydroxytetralin	inactive	(*27*)	inactive	(*27*)	0.0002	(*27*)				
cholesterol	inactive	(*24*)			inactive	(*23*)				
cholesteryl palmitate	inactive	(*24*)								
17α-ethynylestradiol	100[b]	(*26*)			88.8	(*23*)	2.20	(*25*)	125	(*28*)
diethylstilbestrol	234	(*23*)			74.3	(*23*)			250	(*28*)
phenylvinylestradiol	inactive	(*29*)	18	(*29*)						
α-zearalanol (zearanol)	0.026	(*23*)			1.3	(*23*)				
coumestrol	0.024	(*23*)			0.67	(*23*)			0.011	(*28*)
phytosterols	inactive	(*24*)	inactive	(*24*)	inactive	(*24*)				
daidzein	inactive	(*30*)			0.0013	(*23*)				
4-nonylphenol (TG)	0.00036	(*23*)			0.005	(*23*)				
4-octylphenol	inactive	(*23*)			0.003	(*23*)			0.01	(*28*)
di-*n*-butylphthalate	inactive	(*23*)			inactive	(23)				
butylbenzylphthalate	inactive	(*23*)			0.0004	(*23*)				
styrene	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NSD-01	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NSD-08	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NSD-09	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NST-01	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NST-03	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
NST-12	inactive	(*31*)	inactive	(*31*)			inactive	(*32*)		
Bp-3	7.6[b]	(*26*)							105	(*26*)
4-MBC	35.51[b]	(*26*)							79.54	(*26*)
OMC	22.21[b]	(*26*)							61.9	(*26*)
OD-PABA	1.15[b]	(*26*)							51.77	(*26*)
HMS	3.79[b]	(*26*)							36.81	(*26*)
B-MDM	2.01[b]	(*26*)							21.01	(*26*)
bisphenol A	inactive	(*23*)			0.005	(*23*)	13.31	(*25*)	0.0025	(*28*)

[a] RP is the estrogenic potency (%) calculated relative to E2 as reported in the literature. Similarly, RBA is the relative binding affinity (%) compared to E2, ln(PC10) is the natural logarithm of the concentration of the test chemical that exhibits 10% of the transcriptional activity of E2, and RPE is the relative proliferative effect (%) calculated relative to E2. [b] Relative uterotrophic potencies calculated relative to 17α-ethynylestradiol as a positive control.

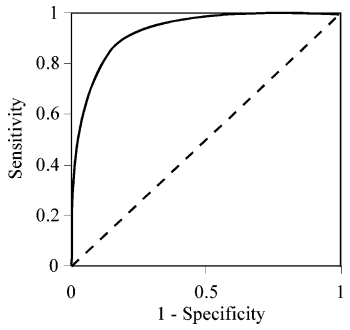

FIGURE 1. Examples of ROC curves. The solid line is typical of an ideal test with high sensitivity and specificity, and the dashed line represents a test unable to discriminate between positive and negative cases.

TABLE 2. Worked Example of a ROC Analysis for the ER Binding Assay[a]

	thresholds							
RBA (%)	0	0.1	18	50	56	60	100	>100
T^+	4	4	3	3	2	1	1	0
sensitivity (T^+/E^+)	1	1	0.75	0.75	0.5	0.25	0.25	0
T^-	0	9	9	10	10	10	10	10
specificity (T^-/E^-)	0	0.9	0.9	1	1	1	1	1
1 - specificity	1	0.1	0.1	0	0	0	0	0

[a] RBA (%) is the relative binding affinity threshold above which a result is considered positive. T^+ and T^- are the number of results over the given threshold for chemicals having shown estrogenicity in vivo or not, respectively. E^+ is the number of chemicals in the dataset that have shown estrogenicity in vivo ($E^+ = 4$), and E^- is the number of chemicals in the dataset that have shown no estrogenicity in vivo ($E^- = 10$).

in terms of chemical structure, and many compounds were of natural origin.

All compounds found to be active in vivo were also found active in vitro and vice versa with inactive compounds except for phenylvinylestradiol. To help with the interpretation of the curves, the details of calculated points for this first example are given in Table 2. The ROC curve (Figure 2) indicates the optimum threshold level to differentiate estrogenic from nonestrogenic compounds on the basis of this dataset would be for a relative binding affinity between 0.1% and 50% depending on the relative importance of sensitivity and specificity desired. Despite the small sample size, the shape of the ROC curve is indicative of a very effective test for estrogenicity.

Reporter gene assays are undertaken with genetically engineered mammalian cells or strains of yeast, with cells transiently or stably transfected with vector DNA sequences for the receptor along with response elements linked to promoter regions for a reporter gene and the reporter gene itself.

YES. Yeast cells *Saccharomyces cerevisiae* stably transfected with the human estrogen receptor α (hERα) gene and expression plasmids with an estrogen responsive element (ERE) are used in this reporter gene assay. The dataset

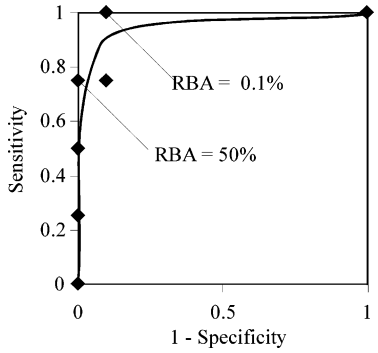

FIGURE 2. ROC curve for the ER binding assay with thresholds for relative binding affinities (RBAs) of 0.1%, which gives the greatest specificity for a sensitivity of 1, and 50%, which gives the greatest sensitivity for a specificity of 1.

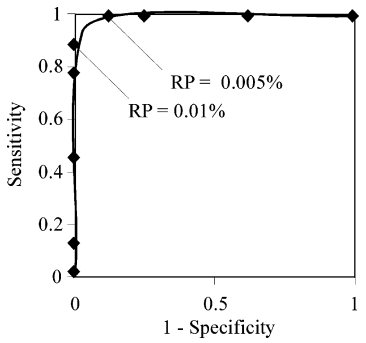

FIGURE 3. ROC curve for YES with thresholds for relative potencies (RPs) of 0.005%, which gives the greatest specificity for a sensitivity of 1, and 0.01%, which gives the greatest sensitivity for a specificity of 1.

included natural and synthetic estrogens and analogues, phytoestrogens and phytosterols, and alkyl phenols. This sample population was positively skewed toward estrogenic compounds, containing nine estrogens (Table 1, second and sixth columns). This positive case prevalence has some influence on the shape of the ROC curve, but it is common to attempt to have roughly the same number of positive and negative cases in an ROC experiment such as to minimize the standard deviation of both sensitivity and specificity. This dataset was however biased toward naturally occurring estrogens.

There were no false negatives in this test, and the yeast assay was more sensitive and less specific than the rodent uterotrophic assay (Table 1). 6-Hydroxytetralin was not found to be active in vivo at oral doses up to 1 (g/kg)/day. 4-Octylphenol, butyl benzyl phthalate (BBP), and bisphenol A were not active in vivo at subcutaneous doses of up to 5 mg. Daidzein was also found to be inactive in vivo. Despite those false positives, the ROC curve (Figure 3) for this limited set of chemicals indicates that YES is a very effective test in predicting short-term acute estrogenicity in vivo. While this may appear as a discrepancy, those false positives were obtained for compounds that exhibited the weakest relative potencies in vitro. ROC analysis can therefore help improve test performance by allowing the objective choice of a threshold over which concordance between in vitro and in vivo results is optimized. In this assay, the threshold level optimizing sensitivity and specificity for our dataset would be for a relative potency of between 0.005% and 0.01% depending on the relative importance of sensitivity and specificity (Figure 3). Additionally, the ROC curves for the ER binding assay and YES are very similar despite relatively different positive case incidence. This illustrates another advantage of ROC analysis, as while it cannot abolish the influence of prevalence, it serves to minimize it (*33*).

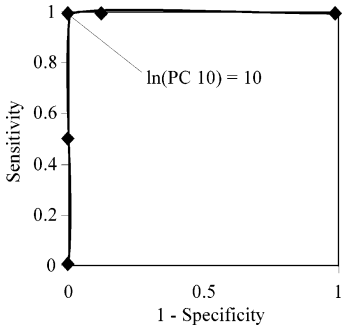

FIGURE 4. ROC curve for the human reporter gene assay with threshold ln(PC10) = 10, which gives both a sensitivity and a specificity of 1.

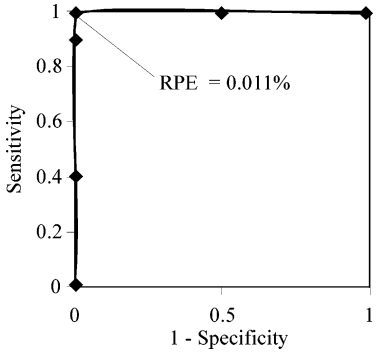

FIGURE 5. ROC curve for the e-screen with a threshold for the relative proliferative effect (RPE) of 0.011%, which gives both a sensitivity and a specificity of 1.

Human Reporter Gene Assay. Mammalian cell lines can also be used in a reporter gene assay; the estrogen responsive cell line used is generally the human breast cancer cell MCF-7. The dataset for this assay was made up of styrene oligomers, and known natural and synthetic estrogens. The styrene oligomers were all inactive in both the rodent uterotrophic assay and reporter gene assay. Although the positive case prevalence was only 2 out of 10, once again the dataset for this assay was not chemically diverse. There was only one false positive result in this ROC test for bisphenol A found to be active in vitro and inactive in vivo (Table 1, second and eighth columns). The best threshold level appears to be for ln(PC10) = 10 for which we obtain maximal sensitivity and specificity values of 1. Therefore, the test would appear perfect as shown by the ROC curve obtained (Figure 4) as over ln-(PC10) = 10 we have perfect concordance between assay results in vitro and those in vivo. However, this may be due to the polarization of the dataset into two extreme groups, known estrogens and inactive styrene oligomers.

e-screen. This technique is based on the measurement of cell proliferation of human-derived breast cancer cells MCF-7 induced through exposure to estrogenic compounds. The dataset for this assay included several sunscreens, a natural estrogen, a synthetic estrogen, a phytoestrogen, and two known xenoestrogens and was therefore strongly biased toward estrogenic compounds. Of the twelve compounds tested with this assay, two compounds gave positive results in vitro and failed to do so in vivo, namely, 4-octylphenol and bisphenol A (Table 1, second and tenth columns). The ROC curve shows that the e-screen would be a perfect test for this set of chemicals. The threshold level for which maximal sensitivity and specificity are obtained would be for an RPE = 0.011% (Figure 5). This again illustrates the threshold effect. Although there were two false positives, all the compounds that exhibited an RPE > 0.011% in the e-screen in vitro assay were active in vivo. Therefore, over a given threshold, we have perfect concordance.

Discussion

Despite the fact that none of the sets of chemicals for each assay contained a large enough number of chemically diverse compounds nor a representative array of the chemicals present in the environment or identified for screening in priority lists, the ROC curves should offer some reassurance with respect to the performance of these bioassays.

Out of the 31 compounds included which were tested by the rodent uterotrophic assay and at least one in vitro assay, none were found inactive in vitro and active in vivo. A larger dataset would probably include some proestrogens, metabolically activated in vivo, while the metabolic capacity of the in vitro bioassays is not well understood. Additional factors that may modulate the uptake and metabolism of xenobiotics include bioaccumulation or interactions involving the induction of plasma binding proteins such as the sex hormone binding globulin (SHBG) or albumin (*34*, *35*). Other false negative results in vitro could arise from estrogenicity mechanisms involving the disruption of the synthesis and metabolism of endogenous hormones or of hormone receptors and/or associated enzymes.

The only false positive in the ER binding assay was for phenylvinylestradiol. This compound was able to bind to the ER but not to elicit an estrogenic response in vivo. This illustrates what would be a likely source of false positive results in a ligand binding assay as these assays cannot discriminate between agonists and antagonists. False positive results in other assay techniques have occurred with compounds such as 4-octylphenol and bisphenol A that are considered known xenoestrogens. However, such discrepancies have already been reported in the literature. Octylphenol and bisphenol A induced a uterotrophic response in vivo 6 h after the last dose was administered, but this response was not maintained 24 h after exposure (*36*), and it has been suggested that bisphenol A exhibits a mechanism of action at the ERα distinct from weak estrogen mimics (*37*). However, bisphenol A was so toxic in the rodent uterotrophic assay at this highest dose that the animals had to be withdrawn from the assay (*23*). This may indicate that the mechanism for short-term acute toxicity for bisphenol A is different from and occurs at a lower exposure level than that for estrogenicity. The design of the in vivo assay and more particularly sample size may have an influence as for ethical reasons one may try to limit the number of animals used. All the false positive results were obtained for compounds that had low relative potencies in YES and the weakest RPE in the e-screen. It may be that a uterotrophic effect was observed but may not have been statistically significant. While validation of the test methods should improve reproducibility of results, this may still be limited by the intrinsic variability of some test variables. The lack of reproducibility of low-dose effects and weak estrogenicity was recently attributed to the natural variability of organ weights (*38*), and this may prove more difficult to address.

There are other limitations with the "truth" criterion, the rodent uterotrophic assay. It is typically a 3 or 4 day assay and detects whether a compound interacts directly or indirectly with the estrogen system. As such, it should still be able to detect estrogenic effects of long-term exposure and exposure at critical life stages, for example in utero, or effects observed in the second or third generation. Estrogenic compounds have a wide range of potencies. Some compounds exhibiting weak estrogenicity in a sensitive in vitro screen may be inactive in vivo. This is not necessarily a discrepancy. Such compounds may be active in vivo and/or in situ in certain circumstances, due to genetic factors, mixture effects, or impaired elimination. However, those potential estrogens may be suitable for a different screening sequence to "true" estrogens. The choice of a truth criterion requires careful consideration as it underlies all subsequent analyses; nonetheless, for the purpose of helping the design of screening sequences, a positive result in the rodent uterotrophic assay, the gold standard for estrogenicity, provides a reliable basis.

The ROC curves obtained show that all the in vitro bioassays were effective at predicting effects in vivo. There is little difference between the curves for different assay techniques. ROC curves could be used to compare the effectiveness of diagnostic tests for a given population by estimating the area under the curves (AUCs) (*14*, *39*). If comparative ROC curves were to be traced, the number of chemicals to include in such a study would need to be fairly large as more cases are needed to demonstrate subtle differences than gross differences (*14*). The array of chemicals selected for an ROC analysis also needs to be representative of the larger population of chemicals to be screened by those tests. The EC priority list (*6*) was drawn on the basis of high production volume and persistence and should be a fair estimation of the chemicals present in the environment, while the USEPA priority list (*7*) is based on exposure potential. Both could be helpful in designing an ROC analysis depending on its intended purpose.

ROC analysis can be used to set or appraise the effectiveness of environmental quality standards by linking concentrations of the compounds of interest in environmental samples to toxicity effects in vivo or even to observed ecological effects in situ, as it allows a quantitative evaluation of the tradeoffs between sensitivity and specificity. This would also be helpful when assessing mixture effects. Other environmental indicators could be assessed in a similar way. For example, a common biomarker used for estrogenicity of river waters is vitellogenin in fish; however, it is unclear whether it indicates any effects at the population level.

In the context of endocrine disruptor screening programs, both the OECD and USEPA screening programs are based on tiered systems. The EPA's Endocrine Disruptors Screening Program (EDSP) is a two-tiered system followed by hazard assessment, which is the ultimate goal of those screening programs. Tier 1 screening includes in vitro and in vivo assays, while tier 2 screening is concerned with life-cycle, critical life stage, and multigenerational in vivo studies (*7*). The OECD Endocrine Disrupters Testing and Assessment (EDTA) Task Force's framework is structured in five levels, starting with prioritization of substances. Level 2 is concerned with in vitro bioassays, level 3 with in vivo assays about a single endocrine mechanism such as the rodent uterotrophic assay, level 4 with in vivo assays about multiple endocrine mechanisms, and level 5 with full or partial life cycle and multigenerational assays (*40*). The present example demonstrated how this approach could be used to assess how effectively in vitro bioassays can predict effects in vivo, but ROC analysis could equally link between other levels of screening. While it offers an objective method of setting thresholds levels, it could also be used to compare test methods and also help determine the combination of screening tests a substance should undergo. Therefore, ROC analysis offers the prospect of nonnegligible potential savings in such a case where tens of thousands of chemical substances need to be screened for several mechanisms of action. The main limitation to this approach is the availability of data. However, as test methods are validated, such data should be made readily available.

ROC analysis can prove as useful a tool for environmental diagnosis and environmental policy as it has been for the biomedical, military, and many other fields. It is an accessible and versatile method whose applications seem to be mainly limited by the availability of suitable data in some instances. In the context of very large screening programs such as in

the case of endocrine disruption, it highlights the benefits of international coordination of research efforts.

Acknowledgments

This study was supported by an Economic & Social Research Council award (PTA-030-2003-00136) and sponsored by the UK Environment Agency. The authors are grateful to Prof. Alan R. Boobis for constructive comments during preparation of the manuscript.

Literature Cited

(1) Krimsky, S. *Hormonal Chaos. The Scientific and Social Origins of the Environmental Endocrine Hypothesis*; The John Hopkins University Press: Baltimore, MD, 2000; pp 1−51

(2) Harvey, P. W., Rush, K. C., Cockburn, A., Eds. *Endocrine and Hormonal Toxicology*; Wiley: Chichester, U.K., 1999; pp. xv−xx.

(3) Jobling, S.; Nolan, M.; Tyler, C. R.; Brighty, G.; Sumpter, J. P. Widespread Sexual Disruption in Wild Fish. *Environ. Sci. Technol.* **1998**, *32*, 2498−2506.

(4) Lai, K. M.; Scrimshaw, M. D.; Lester, J. N. The Effects of Natural and Synthetic Steroid Estrogens in Relation to Their Environmental Occurrence. *Crit. Rev. Toxicol.* **2002**, *32*, 113−132.

(5) Silva, E.; Rajapakse, N.; Kortenkamp, A. Something from "Nothing"—Eight Weak Estrogenic Chemicals Combined at Concentrations Below NOECs Produce Significant Mixture Effects. *Environ. Sci. Technol.* **2002**, *36*, 1751−1756.

(6) European Commission. Towards the Establishment of a Priority List of Substances for Further Evaluation of Their Role in Endocrine Disruption:—Preparation of a Candidate List of Substances as a Basis for Priority-Setting. Final Report, 2000. http://www.europa.eu.int/comm/environment/docum/01262_en.htm#bkh.

(7) U.S. Environmental Protection Agency. Endocrine Disruptor Screening Program Report to Congress, 2000. http://www.epa.gov/scipoly/oscpendo/docs/reporttocongress0800.pdf.

(8) U.S. Environmental Protection Agency. Report to the U. S. House of Representatives Committee on Appropriations on the Status of the Endocrine Disruptor Methods Validation Subcommittee, 2002. http://www.epa.gov/scipoly/oscpendo/docs/edmvs/edmvsstatusreporttocongressfinal.pdf.

(9) Organisation for Economic Co-operation and Development. Detailed Review Paper Appraisal of Test Methods for Sex Hormone Disrupting Chemicals, 2001. http://www.oecd.org/dataoecd/47/21/2074124.pdf.

(10) Ankley, G. T.; Defoe, D. L.; Kahl, M. D.; Jensen, K. M.; Makynen, E. A.; Miracle, A.; Hartig, P.; Gray, L. E.; Cardon, M.; Wilson, V. Evaluation of the Model Anti-Androgen Flutamide for Assessing the Mechanistic Basis of Response. *Environ. Sci. Technol.* **2004**, *38* (23), 6322−6327.

(11) Gomes, R. L.; Scrimshaw, M. D.; Lester, J. N. Determination of Endocrine Disrupters in Sewage Treatment and Receiving Waters. *Trends Anal. Chem.* **2003**, *22*, 697−707.

(12) European Commission. European Workshop on the Impact of Endocrine Disrupters on Human Health and Wildlife (European Commission Report No. Eur 17549), 1997. http://europa.eu.int/comm/environment/endocrine/documents/reports_conclusions_en.htm.

(13) Shine, J. P.; Trapp, C. J.; Coull, B. A. Use of Receiver Operating Characteristic Curves to Evaluate Sediment Quality Guidelines for Metals. *Environ. Toxicol. Chem.* **2003**, *22*, 1642−1648.

(14) Metz, C. E. Basic Principles of Roc Analysis. *Semin. Nucl. Med.* **1978**, *8*, 283−298.

(15) Korach, K. S.; McLachlan, J. A. Techniques for Detection of Estrogenicity. *Environ. Health Perspect.* **1995**, *103*, 5−8.

(16) Shelby, M. D.; Newbold, R. R.; Tully, D. B.; Chae, K.; Davis, V. L. Assessing Environmental Chemicals for Estrogenicity Using a Combination of *in Vitro* and *in Vivo* Assays. *Environ. Health Perspect.* **1996**, *104*, 1296−1300.

(17) Katzenellenbogen, B. S.; Choi, I.; Delage-Mourroux, R.; Ediger, T. R.; Martini, P. G. V.; Montano, M.; Sun, J.; Weis, K.; Katzenellenbogen, J. A. Molecular Mechanisms of Estrogen Action: Selective Ligands and Receptor Pharmacology. *J. Steroid Biochem. Mol. Biol.* **2000**, *74*, 279−285.

(18) Gray, J., L. Earl; Kelce, W. R.; Wiese, T.; Tyl, R.; Gaido, K.; Cook, J.; Klinefelter, G.; Desaulniers, D.; Wilson, E. Endocrine Screening Methods Workshop Report: Detection of Estrogenic and Androgenic Hormonal and Antihormonal Activity for Chemicals That Act Via Receptor or Steroidogenic Enzyme Mechanisms. *Reprod. Toxicol.* **1997**, *11*, 719−750.

(19) Milligan, S. R.; Balasubramanian, A. V.; Kalita, J. C. Relative Potency of Xenobiotic Estrogens in an Acute in Vivo Mammalian Assay. *Environ. Health Perspect.* **1998**, *106*, 23−26.

(20) Owens, W.; Koeter, H. B. W. M. The OECD Program to Validate the Rat Uterotrophic Bioassay: An Overview. *Environ. Health Perspect.* **2003**, *111*, 1527−1529.

(21) Ashby, J. Scientific Issues Associated with the Validation of *in Vitro* and *in Vivo* Methods for Assessing Endocrine Disrupting Chemicals. *Toxicology* **2002**, *181−182*, 389−397.

(22) Beresford, N.; Routledge, E. J.; Harris, C. A.; Sumpter, J. P. Issues Arising When Interpreting Results from an in Vitro Assay for Estrogenic Activity. *Toxicol. Appl. Pharmacol.* **2000**, *162*, 22−33.

(23) Coldham, N. G.; Dave, M.; Sivapathasundaram, S.; McDonnell, D. P.; Connor, C.; Sauer, M. J. Evaluation of a Recombinant Yeast Cell Estrogen Screening Assay. *Environ. Health Perspect.* **1997**, *105*, 734−742.

(24) Baker, V. A.; Hepburn, P. A.; Kennedy, S. J.; Jones, P. A.; Lea, L. J.; Sumpter, J. P.; Ashby, J. Safety Evaluation of Phytosterol Esters. Part 1. Assessment of Oestrogenicity Using a Combination of *in Vivo* and *in Vitro* Assays. *Food Chem. Toxicol.* **1999**, *37*, 13−22.

(25) Yamasaki, K.; Takeyoshi, M.; Yakabe, Y.; Sawaki, M.; Imatanaka, N.; Takatsuki, M. Comparison of Reporter Gene Assay and Immature Rat Uterotrophic Assay of Twenty-Three Chemicals. *Toxicology* **2002**, *170*, 21−30.

(26) Schlumpf, M.; Cotton, B.; Conscience, M.; Haller, V.; Steinmann, B.; Lichtensteiger, W. In Vitro and in Vivo Estrogenicity of Uv Screens. *Environ. Health Perspect.* **2001**, *109*, 239−244.

(27) Elsby, R.; Ashby, J.; Sumpter, J. P.; Brooks, A. N.; Pennie, W. D.; Maggs, J. L.; Lefevre, P. A.; Odum, J.; Beresford, N.; Paton, D.; Park, B. K. Obstacles to the Prediction of Estrogenicity from Chemical Structure: Assay-Mediated Metabolic Transformation and the Apparent Promiscuous Nature of the Estrogen Receptor. *Biochem. Pharmacol.* **2000**, *60*, 1519−1530.

(28) Gutendorf, B.; Westendorf, J. Comparison of an Array of *in Vitro* Assays for the Assessment of the Estrogenic Potential of Natural and Synthetic Estrogens, Phytoestrogens and Xenoestrogens. *Toxicology* **2001**, *166*, 79−89.

(29) Hanson, R. N.; Lee, C. Y.; Friel, C.; Hughes, A.; DeSombre, E. R. Evaluation of 17[Alpha]-E-(Trifluoromethylphenyl)Vinyl Estradiols as Novel Estrogen Receptor Ligands. *Steroids* **2003**, *68*, 143−148.

(30) Jefferson, W. N.; Padilla-Banks, E.; Clark, G.; Newbold, R. R. Assessing Estrogenic Activity of Phytochemicals Using Transcriptional Activation and Immature Mouse Uterotrophic Responses. *J. Chromatogr., B* **2002**, *777*, 179−189.

(31) Date, K.; Ohno, K.; Azuma, Y.; Hirano, S.; Kobayashi, K.; Sakurai, T.; Nobuhara, Y.; Yamada, T. Endocrine-Disrupting Effects of Styrene Oligomers That Migrated from Polystyrene Containers into Food. *Food Chem. Toxicol.* **2002**, *40*, 65−75.

(32) Ohno, K.; Azuma, Y.; Date, K.; Nakano, S.; Kobayashi, T.; Nagao, Y.; Yamada, T. Evaluation of Styrene Oligomers Eluted from Polystyrene for Estrogenicity in Estrogen Receptor Binding Assay, Reporter Gene Assay, and Uterotrophic Assay. *Food Chem. Toxicol.* **2003**, *41*, 131−141.

(33) Collinson, P. Of Bombers, Radiologists, and Cardiologists: Time to ROC. *Heart* **1998**, *80*, 215−217.

(34) Voulvoulis, N.; Scrimshaw, M. D. In *Endocrine Disrupters in Wastewater and Sludge Treatment Processes*; Birkett, J. W., Lester, J. N., Eds.; Lewis Publishers/IWA Publishing: London, 2003; pp 61−72.

(35) Zacharewski, T. *In Vitro* Bioassays for Assessing Estrogenic Substances. *Environ. Sci. Technol.* **1997**, *31*, 613−623.

(36) Laws, S. C.; Carey, S. A.; Ferrell, J. M.; Bodman, G. J.; Cooper, R. L. Estrogenic Activity of Octylphenol, Nonylphenol, Bisphenol A and Methoxychlor in Rats. *Toxicol. Sci.* **2000**, *54*, 154−167.

(37) Gould, J. C.; Leonard, L. S.; Maness, S. C.; Wagner, B. L.; Conner, K.; Zacharewski, T.; Safe, S.; McDonnell, D. P.; Gaido, K. W. Bisphenol A Interacts with the Estrogen Receptor [Alpha] in a Distinct Manner from Estradiol. *Mol. Cell. Endocrinol.* **1998**, *142*, 203-214.

(38) Ashby, J.; Tinwell, H.; Odum, J.; Lefevre, P. Natural Variability and the Influence of Concurrent Control Values on the Detection and Interpretation of Low-Dose or Weak Endocrine Toxicities. *Environ. Health Perspect.* **2004**, *112*, 847−853.

(39) DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837−845.

(40) Organisation for Economic Co-operation and Development. Conceptual Framework for the Testing and Assessment of Endocrine Disrupting Chemicals, 2002. http://www.oecd.org/dataoecd/17/33/23652447.doc.

Received for review September 9, 2004. Revised manuscript received April 15, 2005. Accepted May 3, 2005.

ES048598O