

6.0 *IN VITRO* ER BINDING TEST METHOD PERFORMANCE ASSESSMENT

6.1 Introduction

The ICCVAM Submission Guidelines (ICCVAM, 1999) request that an assessment be conducted of the performance (i.e., accuracy, sensitivity, specificity, positive and negative predictivity, and false positive and false negative rates¹) of the proposed test method with respect to its ability to predict the effect of interest in the reference test method currently accepted by the regulatory agencies and, where feasible, to predict adverse health outcomes in the species of interest (e.g., humans, wildlife). Currently, there are no validated *in vivo* reference test methods developed to specifically assess the ability of a test substance to disrupt endocrine function, and data on endocrine disruption in humans or wildlife are too limited to be used for this purpose. Therefore, the existing *in vitro* ER binding assays were compared against each other with regard to their ability to detect substances capable of binding to the ER. However, this type of analysis of *in vitro* ER binding assays is limited by the lack of multiple test data within and across assays for most of the substances considered, and by the paucity of data for the same substances tested in multiple assays.

Taking these limitations into account, a comparative evaluation was conducted of the relative performance of the 14 *in vitro* ER binding assays considered in this BRD. Both quantitative and qualitative assessments of IC₅₀ and RBA values were conducted. The quantitative assessment was based on the 238 substances (37.3% of the 638 substances in the *in vitro* ER binding assay database) that had been tested in at least two assays (**Appendix E**), and was further limited to individual tests that resulted in an IC₅₀ or RBA value (i.e., the substance was classified as positive). The qualitative assessment was limited to the 100 substances that had been tested in the RUC assay and in at least one of the 13 other *in vitro* ER binding assays, and included substances classified as negative for ER-binding activity.

¹ Accuracy is defined as the proportion of correct outcomes of a method, often used interchangeably with concordance; Sensitivity is defined as the proportion of all positive substances that are correctly classified as positive in a test; Specificity is defined as the proportion of all negative substances that are correctly classified as negative in a test; Positive predictivity is defined as the proportion of correct positive responses among substances testing positive; Negative predictivity is defined as the proportion of correct negative responses among substances testing negative; False positive rate is defined as the proportion of all negative substances that are falsely identified as positive; False negative rate is defined as the proportion of all positive substances that are falsely identified as negative (NIEHS, 1997).

Table 6-1 Number of Substances Tested in Multiple *In Vitro* ER Binding Assays

Number of Assays	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
Number of Substances	403	87	74	23	13	6	7	5	5	4	2	2	3	4	638
% of Substances	63.2	13.6	11.6	3.6	2.0	0.9	1.1	0.8	0.8	0.6	0.3	0.3	0.5	0.6	100

6.2 Quantitative Assessments of Assay Performance

To reduce the extent of skewness in the data prior to conducting the quantitative assessments, the two outcome variables for *in vitro* ER binding assays — the RBA and the IC₅₀ values — were transformed using the natural log. Studies that did not result in an IC₅₀ and/or RBA value were eliminated from consideration. Given the large number of data points for modeling, the general linear models (GLM) used in this analysis are robust, although some skewness may yet exist with the data. To simplify the comparison, each literature citation was considered an independent assessment (designated here as a ‘reference’).

Two-way and three-way analysis of variance models were performed with random effects to estimate the intra-class correlation of substances. A high correlation value indicates that the lnRBA or lnIC₅₀ values are more similar within groups than among groups, where groups can be defined by assay or by reference. Estimates of variance for each model component and intra-class correlation are presented to show which factors (substance, assay, or reference) are responsible for the greatest variation in the lnRBA and lnIC₅₀ values. Due to limitations in the database with regard to the number of substances tested in multiple assays and to the number of independent tests performed for the substance using the same assay, the results of these analyses must be viewed with caution.

Initially, all data representing all substances, assays, and references were considered, and unique data (i.e., substances tested only in a single assay) were excluded from subsequent analyses. For the analysis of lnRBA values, a total of 752 data points representing 211 substances, 14 assays, and 51 references were considered. For the analysis of lnIC₅₀ values, 369 data points representing 119 substances, 13 assays, and 31 references were considered. The lnIC₅₀ and the lnRBA values for 17 -estradiol were omitted from these analyses. The RBA values for 17 -

estradiol are uninformative because they are arbitrarily set at 100% in all assays in which this substance is used as the reference estrogen. The IC₅₀ values for 17 β -estradiol represent the largest collection of IC₅₀ data for a single substance and were evaluated independently to avoid potentially biasing the quantitative analysis.

6.2.1 Measures of Intra-Class Correlation

The intra-class correlation, r_I , measures the percentage of variation in y , the outcome variable, explained by a given component or set of components. The model is $y = \text{substance} + \text{assay} + \text{reference}$. **Table 6-2** contains the components of variance for each variable adjusted for the other two variables. Interpretation of this analysis is limited to factors that impact on performance; factors that impact on assay reliability are discussed in **Section 7**.

From this analysis, it appears that the lnRBA or lnIC₅₀ values for a specific substance were generally consistent irrespective of which assay was used or which laboratory conducted the study. The greatest variation in lnRBA or lnIC₅₀ values was found between substances (i.e., the most important parameter was the intrinsic ER binding property of the substance). The greater contribution of substances to the overall variance is not surprising considering the seven orders of magnitude range in reported IC₅₀, and thus RBA, values.

6.2.2 Evaluation of Substances Tested in Nine or More *In Vitro* ER Binding Assays

In this analysis, the variances for the RBA values of the 12 substances that had been tested in at least nine of the 14 *in vitro* ER binding assays were determined. Although 14 substances (excluding 17 β -estradiol) had been tested in at least ten *in vitro* ER binding assays (**Section 5**), only those substances that elicited a positive response in at least one experiment in each assay could be used in this analysis. The variances and sample sizes for these 12 substances are provided in **Table 6-3**, ranked in descending order according to the median RBA value based on all test data. Only assays for which variances could be calculated are included, and most of these variances were based on three or four values only. Due to the lack of sufficient data, a corresponding analysis of IC₅₀ values was not conducted.

Table 6-2 Components of Variance for Each Variable Adjusted for the Other Two Variables – Performance Assessment

	Outcome, y (% variation)	
	lnRBA	lnIC ₅₀
Var(substance)	8.34	8.49
Var(assay)	0.38	0.34
Var(reference)	1.40	2.01
Var(error)	1.75	2.44
Corr (y _{ijk} , y _{ij'k'})*	0.70	0.64
Corr (y _{ijk} , y _{ijk'})**	0.73	0.67
Corr (y _{ijk} , y _{ij'k'} ***)	0.82	0.79

*A high correlation was found for the lnRBA values within substances using any assay or reference (i.e., the lnRBA values are more correlated within than across substances). A slightly lower correlation was found when lnIC₅₀ values were used. The high correlation for the lnRBA values suggests that the RBA of a specific substance to the ER did not vary much among the different binding assays.

**This correlation suggests that the test substances responded similarly in an assay irrespective of the laboratory in which the test was conducted. Variation within laboratories is slightly less than the variation across laboratories.

*** A high correlation was found for substances tested in the same laboratories (i.e., references) but using different assays.

A large p value (p1 or p2) identifies those substances, such as zearalenone, estriol, estrone, diethylstilbestrol (DES), 2,2-bis(*p*-hydroxyphenyl)-1,1,1-trichloroethane (HPTE), bisphenol A, and kepone, with the least amount of variability in their lnRBA values. In contrast, the p1 values of coumestrol and tamoxifen are below 0.05, indicating that significant variability exists across assays irrespective of the laboratory in which the tests were performed. A possible explanation for the variability with coumestrol, a phytoestrogen, is its ~1.5-log greater binding affinity to the ER protein compared to the ER protein (**Appendix D**). No explanation can be provided for the significant variability in lnRBA values for tamoxifen. Values for p2 could not be calculated in every case since there were too few assays or references that could be used in the analysis. A significant p2 value was not found for any substance suggesting that there was not significant variability due to the reference (i.e. laboratories in which the substance was tested).

Another approach to evaluating the variability across assays for a substance is to fit a two-way model, where $y = \text{assay} + \text{reference}$. In this analysis (**Table 6-4**), adjustment is made for inter-reference variation in lnRBA so that only those assays used twice or more in two or more

Table 6-3 Variance of lnRBA Values by Substance and Assay – Performance Assessment^a

Substance ^b (CASRN)	Median ^c RBA	#of Obs/ # Assays	hER α ^d	hER α - FP ^d	hER β ^d	MCF-7 cytosol ^d	MUC ^d	RUC ^d	p1*	p2**
4-Hydroxy- tamoxifen (68047-06-3)	168	18/13	0.28 (3)	1.82 (3)					0.08	0.15
DES (56-53-1)	127	38/14	0.99 (3)	0.45 (4)			0.60 (7)	3.42 (11)	0.15	0.99
Estrone (53-16-7)	45	18/13				2.40 (3)		0.98 (4)	0.73	na ^e
Estriol (50-27-1)	15.8	16/12				2.42 (4)			0.53	0.64
Zearalenone (17924-92-4)	15.0	11/9	All n \leq 2						0.42	na
Tamoxifen (10540-29-1)	5.0	21/14	0.44 (3)					2.01 (4)	0.02	0.10
Coumestrol (479-13-0)	3.1	15/11	0.79 (3)						0.02	0.25
HPTE (2971-36-0)	1.45	12/10						1.53 (3)	0.82	na
Genistein (446-72-0)	1.30	18/11	1.07 (4)		0.97 (3)				0.11	0.18
Bisphenol A (80-05-7)	0.031	22/14	1.36 (3)					1.25 (5)	0.53	0.60
<i>o,p'</i> -DDT (789-02-6)	0.038	15/10						1.72 (5)	0.20	na
Kepone (143-50-0)	0.027	11/9						1.39 (3)	0.60	na

^aOnly assays where a variance could be calculated for at least one of the 12 substances are listed. The variance for a particular assay could be calculated only if a particular substance was tested three or more times in that assay; empty cells indicate insufficient data to calculate a variance. The p values could be calculated only if there were two observations from at least three or more assays; a missing p-value indicates insufficient data.

^bSubstances that had been tested in at least 9 of the 14 *in vitro* ER binding assays; DES = diethylstilbestrol; *o,p'*-DDT = *o,p'*-dichlorodiphenyltrichloroethane; HPTE = 2,2-Bis(*p*-hydroxyphenyl)-1,1,1,-trichloroethane.

^cThe median RBA value across assays, based on positive test data.

^dThe numbers in parenthesis indicate the numbers of replicate tests.

^ena = No p value could be calculated since there was either no values or only one value per assay x response combination.

*p1 tests whether there is a significant difference among all assays used; unadjusted for references.

**p2 tests whether there is a significant difference among all assays used; adjusted for references.

laboratories (references) are considered. Results are presented in descending order according to the median RBA value across assays, based on all positive test data, for each of the 12 substances. The components of the variance for each variable are adjusted for the other variable. Due to the lack of sufficient data, a corresponding analysis of IC₅₀ values was not conducted.

Table 6-4 Variance for Y=lnRBA Values

Substance ^a (CASRN)	Median RBA ^b	N ^c	n/n ^c	var(assay)	var(ref)	var(error)	r _i ^d (assay)
4-Hydroxy-tamoxifen (68047-06-3)	168	18	13/8	0.66	1.58	0.17	0.27
DES (56-53-1)	127	38	14/8	≤0.001	6.37	0.35	~0 ^e
Estrone (53-16-7)	45	18	13/7	0.25	2.88	0	0.08
Estriol (50-27-1)	15.8	16	12/7	0.096	4.54	0.49	0.001
Zearalenone (17924-92-4)	15.0	11	9/6	0.27	Too few references	0.44	0.38
Tamoxifen (10540-29-1)	5.0	21	14/8	0.53	1.91	0.08	0.21
Coumestrol (479-13-0)	3.1	15	11/7	0.49	0.22	0.43	0.43
HPTE (2971-36-0)	1.45	12	10/6	1.14	2.34	0	0.33
Genistein (446-72-0)	1.30	18	11/7	1.41	≤0.001	1.23	0.53
<i>o,p'</i> -DDT (789-02-6)	0.038	15	10/4	2.89	2.90	0	0.50
Bisphenol A (80-05-7)	0.031	22	14/8	≤0.001	≤0.001	2.64	~0
Kepone (143-50-0)	0.027	11	9/6	0.84	1.93	0	0.30

^aSubstances that had been tested in at least nine of the 14 *in vitro* ER binding assays; DES = diethylstilbestrol; *o,p'*-DDT = *o,p'*-dichlorodiphenyltrichloroethane; HPTE = 2,2-Bis(*p*-hydroxyphenyl)-1,1,1-trichloroethane

^bThe median RBA value across assays, based on positive test data.

^cN is the total number of values available; n is the number of assays used to test that substance; and n' is the number of assays that can be adjusted for the effect of reference to generate the data in this table.

^dr_i, the intra-class correlation, measures the percentage of variation in y, the outcome variable, explained by a given component or set of components

^er_i= 0 when each RBA value is derived from a different assay x reference combination

As demonstrated by the relatively small intra-class correlation values, the lnRBA values are very similar across assays for estriol and estrone, and not quite as similar across assays for tamoxifen, HPTE, kepone, and 4-hydroxytamoxifen. The relatively large intra-class correlation values for genistein, coumestrol, *o,p'*-DDT and zearalenone suggest that these substances respond differently in the various assays. The explanation for the increased variability associated with genistein and coumestrol, both of which are phytoestrogens, might be their ~1.5-log greater binding affinity to the ER α protein compared to the ER β protein used in other assays. No explanation can be provided for the increased variability in lnRBA values associated with zearalenone and *o,p'*-DDT. However, the lack of an obvious relationship between the magnitude of the median RBA value for a substance and its intra-class correlation value suggests that the increased variability across assays for some substances is not a reflection of its binding activity. This analysis is affected to a great extent by the fact that so few assays were used within the same reference.

6.2.3 Variability in lnIC₅₀ and lnRBA Values for Selected Substances

Another approach for assessing the variability between substances is to evaluate the standard deviation of the lnRBA and lnIC₅₀ values of the 12 substances tested in at least nine of the 14 *in vitro* ER binding assays. These data are tabulated along with the corresponding median RBA values across assays in **Table 6-5**. The standard deviations were visually compared to determine which substances demonstrate more variability than others if the effects of assay and laboratory, which appear to be relatively small, are ignored. The overall variability presented in **Table 6-5** and the variability across and within assays shown in **Table 6-4** should be considered together.

The least amount of variation in binding affinity (based on assessing both lnRBA and lnIC₅₀ values) occurred for zearalenone, while the greatest variations (twice the lowest value) were observed for coumestrol, *o,p'*-DDT, and DES. Among the other substances, the variability in binding affinity was relatively similar among the different assays. Increased variability in the lnRBA and lnIC₅₀ values for coumestrol may be related to its much higher binding affinity for the purified proteins, especially ER α , compared to the cytosolic receptors (**Appendix D**).

Table 6-5 Variability in Standard Deviations for lnRBA and lnIC₅₀ Values For Selected Substances

Substance ^a (CASRN)	Median ^b RBA	# of Assays	lnRBA		lnIC ₅₀	
			Standard Deviation	N ^c	Standard Deviation	N ^c
4-Hydroxy- tamoxifen (68047-06-3)	168	13	1.36	18	1.68	10
DES (56-53-1)	127	14	2.01	38	3.20	26
Estrone (53-16-7)	45	13	1.49	18	1.57	8
Estriol (50-27-1)	15.8	12	1.36	16	0.89	6
Zearalenone (17924-92-4)	15.0	9	0.84	11	0.76	8
Tamoxifen (10540-29-1)	5.0	14	1.91	21	1.68	13
Coumestrol (479-13-0)	3.1	11	2.30	15	2.51	9
HPTE (2971-36-0)	1.45	10	1.15	12	1.14	10
Genistein (446-72-0)	1.30	11	1.74	18	1.64	12
<i>o,p'</i> -DDT (789-02-6)	0.038	10	2.27	15	1.87,	12
Bisphenol A (80-05-7)	0.031	14	1.63	22	1.54	15
Kepon (143-50-0)	0.027	9	1.37	11	1.07	8

^aSubstances that had been tested in at least 9 of the 14 *in vitro* ER binding assays; DES =diethylstilbestrol; *o,p'*-DDT=*o,p'*-dichlorodiphenyltrichloroethane; HPTE=(2,2-Bis(*p*-hydroxyphenyl)-1,1,1,-trichloroethane.

^bThe median RBA value across assays, based on positive test data.

^cN indicates the number of RBA or IC₅₀ values used in the analysis.

6.3 Qualitative Assessment of *In Vitro* ER Binding Assay Performance

A qualitative comparative assessment of assay performance considered the relative ability of the 14 *in vitro* ER binding assays to identify substances with relatively weak ER binding affinities and to obtain higher RBA values for the same set of substances. In conducting this assessment, it was assumed that all positive study results and all negative results for studies in which the highest dose tested was at least 100 µM were correct, for that assay. The 100 µM dose level criterion for negative studies was used to ensure that the protocol (in terms of test substance dose

levels) was minimally adequate for detecting weak positive responses. Thus, a positive assay reflects the intrinsic ability of the test substance to bind to the ER while a negative assay reflects difference in assay sensitivity rather than differences in the experimental protocol.

Due to the RUC assay having the largest database, this assay was used as the standard to compare with the performance of each of the 13 other *in vitro* ER binding assays. To conduct this assessment, the median RBA value was calculated for any substance tested positive in two or more tests using the same assay; otherwise the RBA value for a single positive test was used for that assay. Next, the resulting single or median RBA value for each substance in each assay was classified into one of seven RBA activity categories -- 100, from <100 to 10, from <10 to 1, from <1 to 0.1, from <0.1 to 0.01, from <0.01 to 0.001, and <0.001. This classification scheme categorizes the range of RBA values into the seven orders of magnitude reported for ER binding substances (**Appendix D**). Substances that tested negative (i.e., no RBA value could be calculated) were classified as negative for that test. In situations where both positive and negative test results were obtained for the same substance using the same *in vitro* ER binding assay, the substance was classified as equivocal within the RBA value category for the positive assay(s). The RBA value category obtained for a substance tested in any *in vitro* ER binding assay other than the RUC assay was then compared and classified as higher, the same, lower, or negative in relation to the RBA value category obtained for that substance in the RUC assay. The results were then inspected to identify assays that appeared to have performed (1) better than, (2) as well as, or (3) not as well as the RUC assay. Improved performance for an assay would be demonstrated by a shift in the RBA values for substances tested in common to higher RBA value categories and to having fewer negative calls, compared to the RUC assay. Equal performance would be demonstrated by both the RUC and the assay being considered having the same RBA value categories for the majority of substances tested in common. Decreased performance for an assay would be demonstrated by a shift in the RBA values tested in common to lower RBA value categories and to having more negative calls, compared to the RUC assay. The results of this approach are summarized in **Table 6-6**.

This qualitative assessment is confounded by a number of limitations, including:

- The lack of multiple test data within an assay for the majority of the substances considered;

- The lack of a common set of substance to compare across all assays;
- The limited number of substances tested in common between the RUC and any other assay;
- The assumption that each test was conducted appropriately and that all test results were accurate for that assay;
- The arbitrariness of the RBA value categories and the possible adverse effect substances with RUC RBA values near the boundary between any two RBA value categories have on the assessment; and
- The inherent complexity added to an assessment when equivocal test substances (i.e., those with multiple, discordant test results) are classified as positive only.

Despite the limitations, the assessment suggests that:

- The hER , hER -FP, hER , and rER assays performed better than the RUC assay, as demonstrated by a shift among the substances tested toward higher category RBA values.
- The GST-ERdef assays, except for GST-rtERdef, did not perform as well as the RUC assay, as demonstrated by a shift among the substances tested toward lower category RBA values and more substances classified as negative. Many of the negative tests were for substances classified as equivocal in the RUC assay and tested only once in the GST-ERdef assays, potentially limiting the validity of this conclusion. The GST-rtERdef assay performed as well as the human and rat ER / assays.
- The MCF-7 cell assay did not perform as well as the RUC assay (increased numbers of substances with lower RBA value categories/negative results), while the MCF-7 cytosol assay performed about the same as the RUC assay.
- For the two other animal based test methods, the MUC assay performed better than and the RBC not as well as the RUC assay.

Table 6-6 Qualitative Assessment of the Ability of Different ER Binding Assays to Detect Substances with Different Relative Binding Affinities (RBA Values) Compared to the RUC Assay

Assay	Result	RBA Value Range								Totals
		≥100	<100-10	<10-1	<1-0.1	<0.1-0.01	<0.01-0.001	<0.001	Negative	
RUC (97)^a	+	6	17	6	13	16	8	3	0	
	+/-	0	0	0	1	1	4	7	0	
	-								15	
hERα (48)	Higher	-	2	1	4	7	3	2	2	21
	same	3	4	1	4	6	0	0	2	20
	lower	1	1	0	3	0	0	-	-	5
	negative	0	0	0	0	1	0	1	-	2
hERα-FP (24)	Higher	-	0	1	2	4	2	1	1	11
	same	1	1	1	1	2	0	0	2	8
	lower	1	1	0	0	1	0	-	-	3
	negative	0	0	0	0	0	1	1	-	2
hERβ (32)	Higher	-	2	1	4	7	3	2	2	19
	same	3	4	1	4	6	0	0	2	9
	lower	1	1	0	3	0	0	-	-	4
	negative	0	0	0	0	1	0	0	-	0
rERβ (24)	Higher	-	2	1	3	2	1	0	0	9
	same	3	4	1	2	0	0	0	2	12
	lower	0	0	0	0	0	0	-	-	0
	negative	0	0	0	0	1	0	1	-	2
GST- aERdef (28)	Higher	-	0	1	3	2	1	0	1	8
	same	3	5	1	1	1	0	0	1	12
	lower	0	0	0	1	0	0	-	-	1
	negative	0	0	0	0	1	2	4	-	7
GST- cERdef (27)	Higher	-	0	1	2	0	0	0	1	4
	same	3	5	1	2	3	1	0	1	16
	lower	0	0	0	0	1	0	-	-	1
	negative	0	0	0	0	0	2	4	-	6

Assay	Result	RBA Value Range								Totals
		≥100	<100-10	<10-1	<1-0.1	<0.1-0.01	<0.01-0.001	<0.001	Negative	
GST-hERαdef (28)	Higher	-	0	1	3	0	0	0	0	4
	same	2	4	0	2	1	0	0	2	11
	lower	1	1	1	0	2	0	-	-	5
	negative	0	0	0	0	1	3	4	-	8
GST-mERαdef (27)	Higher	-	0	1	2	0	0	0	0	3
	same	2	5	0	2	1	0	0	2	12
	lower	1	0	1	0	3	0	-	-	5
	negative	0	0	0	0	0	3	4	-	7
GST-rtERdef (29)	Higher	-	2	1	4	2	2	3	2	16
	same	3	3	0	1	2	0	0	1	10
	lower	0	0	1	0	0	0	-	-	1
	negative	0	0	0	0	0	1	1	-	2
MCF-7 cells (21)	Higher	-	0	0	0	0	1	0	2	3
	same	1	2	1	1	2	0	0	0	7
	lower	3	5	2	0	1	0	-	-	11
	negative	0	0	0	0	0	0	0	-	0
MCF-7 cytosol (31)	Higher	-	0	2	3	1	2	0	0	8
	same	4	10	3	0	1	0	0	0	18
	lower	0	2	0	1	1	0	-	-	4
	negative	0	0	0	0	0	1	0	-	1
MUC (24)	Higher	-	3	0	2	2	3	1	0	11
	same	1	1	1	1	4	0	0	1	9
	lower	1	0	0	2	0	0	-	-	3
	negative	0	0	0	0	1	0	0	-	1
RBC (22)	Higher	-	1	0	1	0	0	0	0	2
	same	3	3	2	0	0	1	0	0	9
	lower	1	0	1	0	3	0	-	-	7
	negative	0	0	0	0	1	3	2	-	4

^aNumber of substances.

Assessment based on substances tested in the RUC assay and at least one other *in vitro* ER binding assay. Data for the RUC assay entered as the number of positive (+), equivocal (+/-) (i.e., the substance was tested in more than one test with both positive and negative results obtained), and negative (-) calls for substances tested in that assay. Higher, the same, lower, and negative results signifies the occurrence of a higher, the same, lower, or negative RBA values compared to the corresponding RBA value obtained in the RUC assay for the same substance. Negative test method results in which the highest dose tested was <100 μ M were not included in this assessment.

6.4 Performance of *In Vitro* ER Binding Assays

The *in vitro* ER binding assays that are the most useful as a screen for endocrine disruptors are those that are the most sensitive (i.e., have the greatest ability to detect weak ER-binding substances) and the most reliable (i.e., exhibit the lowest variance) (see **Section 7**). In addition, it might be anticipated that those assays that use ER derived from the species of interest (e.g., human for predicting human-related effects, wildlife species for predicting effects in wildlife) might be the most informative. Finally, when taking animal welfare and human health and safety issues into consideration, assays that do not use ER obtained from experimental animals or ones that do not use radioactivity, respectively, might be of the greatest utility.

The results of the quantitative and qualitative assessments of the performance of the 14 *in vitro* ER binding assays evaluated in this BRD, as well as the results of an assessment of the utility (source of ER, absence of animal use, absence of the use of radioactivity) of the various assays, are summarized in **Table 6-7**. Based on these assessments, the hER_α, hER_α-FP, hER_β, and GST-rtERdef assays appear to offer the greatest overall performance and utility as screening assays. The receptor used in the GST-rtERdef assay is derived from the rainbow trout and thus might be less relevant for the screening of substances that might affect endocrine function in humans. However, this assay might have greater utility in screening for ED substances that might impact wildlife. The relative utility of ER_α versus ER_β assays in a screening paradigm needs further consideration. Among the substances tested in both the assays, 55% produced a higher RBA value in a hER_α assay, while 24% produced a higher RBA value in a hER_β assay. This suggests that a hER_α assay might perform better in a screening battery. As another consideration, the ER_α protein predominates in the uterus, while the ER_β protein is predominant in the prostate gland (Kuiper et al., 1997). Thus, inclusion of both types of estrogen receptors in a screening battery might be advantageous. However, among the 82 substances tested in common between the two assays, only two substances were discordant (i.e., one test substance was positive in a hER_α assay but negative in a hER_β assay, and vice-versa), suggesting that either assay would perform equally well in a screening battery.

Table 6-7 Summary of *In Vitro* ER Binding Assay Performance

Assay	Quantitative Performance ^a	Qualitative Performance ^b	Use of Experimental Animals ^c	ER from Species of Interest ^d	Non-radioactive Technology ^e
RUC	0				
hER	0	+	+	+	
hER -FP	0	+	+	+	+
hER	0	+	+	+	
rER	0	+	+		
GST-aERdef	0	-	+		
GST-cERdef	0	-	+		
GST-hER def	0	-	+	+	
GST-mER def	0	-	+		
GST-rtERdef	0	+	+	+	
MCF-7 cells	0	-	+		
MCF-7 cytosol	0	0	+		
MUC	0	+			
RBC	0	-			

^aThe quantitative assessment did not convincingly indicate that any single assay performed better than any other assay

^bThe RUC assay was used as the standard assay in the qualitative assessment; + = assays with improved performance; 0 = assays with similar performance; - = assays with lower performance than the RUC assay.

^cUtility (+) based on the lack of need for experimental animals.

^dUtility (+) based on the use of ER from a species of direct interest (i.e., human ER for human health, a wildlife species for ecological effects).

^eUtility (+) based on the use of non-radioactive technology.

6.5 General Strengths and Limitations of *In Vitro* ER Binding Assays

Competitive binding assays indicate whether a substance can interact with the target receptor by its ability to displace the natural ligand. These assays do not provide sufficient evidence to conclude that a substance is an agonist or an antagonist, or take into consideration other mechanisms of action that may lead to endocrine disruption (Zacharewski, 1998). However, *in vitro* binding assays can be important components of a battery of tests and are suitable for screening, because they:

- Are cost-effective;
- Are rapid and relatively easy to perform;

- Are based on a easily quantitated, well-elucidated mechanism of action (i.e., binding to a specific protein);
- Are sensitive (50 fmol ER/mg protein can be detected);
- Can be performed using small amounts of test substances;
- Can be used to test multiple substances simultaneously; and
- Can be easily standardized among laboratories.

These assays have limitations also, including:

- Inability to distinguish agonists from antagonists; and
- Potential generation of false positive and false negative results.

In terms of false positive results, the substance might disrupt the binding of the radioactive ligand to the ER by deactivating the receptor or decrease binding via noncompetitive inhibition (Kupfer, 1988). The latter might occur at high concentrations of the test substance. For false negative results, the accurate measurement of rapidly dissociating, low affinity ligands can be difficult because the bound ER and ligand are not in equilibrium when the unbound ligand is washed away from the receptor. Under these conditions, low affinity ligands are more likely to dissociate from the ER. This dissociation is a concern when the receptor or ligand is bound to a solid support such as charcoal that is used in traditional competitive ER binding assays (National Academy of Sciences, 1999). Assays that use FP to assess ER changes would not be affected by this concern. Other mechanisms for obtaining a false negative response include metabolic activation of the test substance to an active intermediate, which subsequently binds to the ER, incomplete solubility in the assay buffer, or incompatibility with assay conditions. Because traditional ER binding assays do not include the enzymes and co-factors required for metabolic activation, some potential ER binding substances will be missed. A possible solution to this limitation is to develop *in vitro* ER binding assays that include a metabolic activation system, as has been conducted in some ER TA assays (Charles et al., 2000; Sumida et al., 2001).

6.6 Conclusions and Recommendations

Although a large number of substances have been tested in *in vitro* ER binding assays, relatively few substances have been tested more than once in the same assay or in multiple assays.

Furthermore, as the primary focus of many of the investigations using *in vitro* ER binding assays has been at understanding mechanisms of binding and transcriptional activation and not at identifying substances with ER binding activity, much of the published data are of limited value in terms of an analysis of performance. Although these limitations weaken the validity of any assessment of *in vitro* ER binding assays, some general conclusions can be made.

The quantitative assessment of lnRBA and lnIC₅₀ values determined that the effect of substances on the variation in RBA and IC₅₀ values was much greater than the effect of assay type, and that significant differences in performance among the different *in vitro* ER binding assays were not present. One limitation of the quantitative assessment was that this approach does not consider situations in which a substance was classified as negative and positive in different tests using the same assay. The qualitative assessment considered whether RBA values (single or median) obtained for substances tested in each of 13 assays were within the same log range as the corresponding values obtained for the same substances in the RUC assay, and whether substances reported as positive or negative in the RUC assay were classified as negative or positive, respectively, in other assays. The RUC assay was selected as the assay for comparison because it had the largest database with respect to the number of substances tested and the number of laboratories using the procedure. The explicit assumption in this assessment was that an assay would perform as well as or better than the RUC assay if it demonstrated similar or higher RBA values and had the same or fewer negative calls for the same set of substances, respectively. Using this approach, the hER /hER -FP, hER /rER , GST-rtERdef, and the MUC assays appear to have performed better than the RUC assay, while the MCF-7 cytosol assay appears to have performed about as well as the RUC assay. The remaining eight assays did not perform as well as the RUC assay but this may reflect the level of usage and the types of substances tested rather than a lack of performance. Similar to the quantitative assessment, this approach is limited by the lack of multiple test data within an assay for most of the substances considered, and by the lack of a common substance database to compare across all assays. The assessment also assumes that each test was conducted appropriately and that the test results were accurate.

Taking into account the available *in vitro* ER binding assay database and the various quantitative and qualitative assessments conducted on the 14 *in vitro* ER binding assays considered in this BRD, the following recommendations can be made in regard to the use of such assays as screening test methods within a battery of Tier 1 endocrine disruptor tests.

- Based on a consideration of such factors as relative performance, elimination of animal use, the use of the ER from the species of interest, and the use of alternatives to radioactive substances, the hER , hER -FP, and hER assays should have the highest priority for validation as screening assays for human health-related issues, while the GST-rtERdef assay might be preferred when screening for substances that pose a hazard to wildlife. Due to an inability to conduct an adequate assessment of assay reliability (see **Section 7**), reliability was not considered in making these recommendations. However, it might be expected that assays which use semi-purified or purified ER proteins would be more reliable than those based on extracts of ER from animal tissues.
- In conducting future validation studies with these assays, the RUC assay should be used as the reference test method. The RUC assay is currently undergoing validation efforts sponsored by the U.S. EPA and the resulting performance and reliability information could be used to establish minimal performance standards for other assays.
- Formal validation studies should be conducted using appropriate substances covering the range of expected RBA values to adequately demonstrate the performance characteristics of the *in vitro* ER binding assays recommended as possible screening assays. A list of potential test substances for use in such a validation effort is provided in **Section 12**.
- There is little information about the ER binding activity of metabolites of xenobiotics and it is not clear whether metabolic activation needs to be included in *in vitro* ER binding test methods used as screening assay. This issue should be considered prior to the implementation of future validation studies.

[This page intentionally left blank]