Progress in the International Harmonization of Estrogen Screens

James W. Owens

Procter & Gamble

Good morning – I'm Dr. Willie Owens of Procter & Gamble, and I serve on both the OECD and the USEPA committees standardizing and validating endocrine screens and tests. My presentation is to review the rapid progress in the last 3 years on the standardization and validation of screens for estrogen activity, particularly within the OECD.

To successfully complete this effort, we must continue international cooperation and harmonization of estrogen screening and testing – Without this consensus for a common approach, we have fragmented pieces. Rather than an efficient and economical international effort, we have different assays run on the same compound – effectively wasting resources – and generating different results and interpretations.

To state the core problem: A large number of compounds need to be screened, maybe as many as 90,000 or more substances are in commerce.

National regulatory agencies can either take different approaches, leading to expensive duplication and potentially conflicting data, or agree on a common approach, sharing the work and data in order to have the greatest efficiency.

The solution proposed by national and international workshops is an efficient, stepwise framework arranged in the following tiers:

1. Review available data – constructing a common, accessible database.
2. Assess structure activity relationships to identify compounds needing screening.
3. Employ *in vitro* mechanistic screens to further prioritize compounds.
4. Employ *in vivo* screens, such as the uterotrophic bioassay – to identify those compounds warranting testing for adverse effects.
5. Conduct the necessary tests for adverse effects.

Validation calls for methods to have clear rationales, so let us review these.

- SARs reduce the large universe of compounds to a workable subset
- *In vitro* screens are rapid, economical, and provide sensitivity – further reducing the number of compounds
- *In vivo* screens incorporate toxicokinetics and thus provide a relevant prediction, to give clear candidates for testing

Each tier contributes its capabilities to reduce the number of unnecessary chemicals without false negatives and identifies the need and priorities for testing.

To validate the screens and the framework - the characteristics of estrogens at the testing phase must be clear.

In this slide are several functional reproductive and developmental endpoints linked to doses of potent estrogens such as 17beta-estradiol, DES, and EE in reproductive studies. All are easily measured in current reproductive protocols.

- mating and sexual behavior, time to mating
- ♂ and ♀ fertility and fecundity (litter size and survival)
- gestation length, implantation loss, premature delivery
- maternal lactational behaviors
- decreased pup weights and survival (pnd 1, 4, 10)

Other estrogen-sensitive endpoints in multiple generation protocols are listed here.

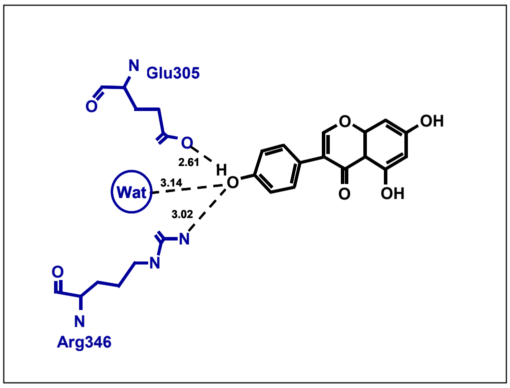
- ♂ and ♀ gonad development (morphology, weight, histopathology)
- reproductive tract and accessory sex organ development (morphology, weight, histopathology)
- sexual developmental benchmarks: vaginal opening, preputial separation, first estrus
- estrous cyclicity
- epididymal sperm numbers and morphology; motility testicular spermatid head counts; daily sperm productionTogether, the two slides constitute a profile or fingerprint for estrogens.
- Weak estrogen agonists expected to first trigger one or more of the endpoints on this second slide,
- With the sensitive endpoints emerging first – as noted later, vaginal opening may be the most sensitive endpoint, and
- The others emerging with increasing doses
- and recognizing that other toxicities may be observed before any estrogen activity is found.

There are a limited number of estrogens where a sufficient body of data exists for all of the proposed screens and tests – these chemicals then become the references to validate each step and the overall framework. The most important data are recent multiple generation test data incorporating the endpoints in the previous slides. These include 17β-estradiol, ethinyl estradiol, genistein, nonylphenol, octylphenol, methoxychlor, and bisphenol A.The basic principles for using SARs are– keep the approach as simple as possible, do not overcomplicate with heavy computations for all chemicals. Consider the use of subtiers – use structure alerts first as filters so that computational complexity should come last. SARs do not have to be perfect – they should only provide direction and suggest priorities and they need to be constructed to avoid false negatives.

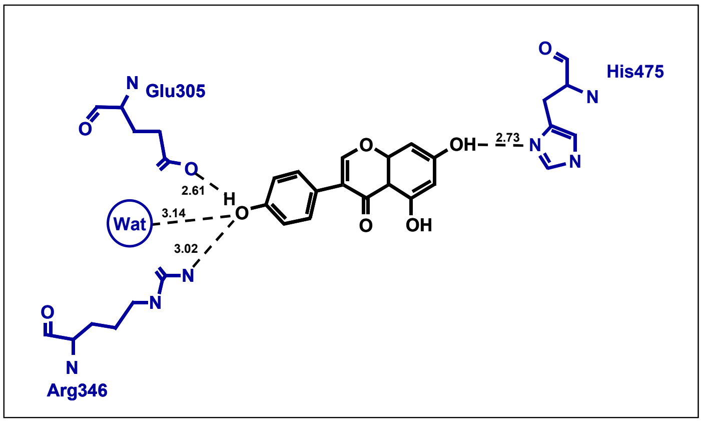
For example, validation calls for a clear scientific rationale and model consistent with the intended regulatory use. In this case, the estrogen ligand must fit into a spatially limited cavity or pocket in the estrogen receptor. The actual upper limit for an agonist appears to be about 350 cubic angstroms and they are organic carbon chemicals. Thus, we have key determinants to construct an initial filter.

This means that molecular weight cut-offs, not too limited, have a clear scientific rationale as a starting point for the SAR. And the effect of such simple filters can be dramatic, cutting the chemicals considered by about half using a 95-1000 MW and organic criteria.

There are also scientific rationales for SAR structural alerts. First, the receptor has two amino acids and a water molecule that align the ligand in the receptor pocket using donating and accepting bonds with a hydroxyl group. Here genistein is the ligand. This interaction appears to be common for all ligands.



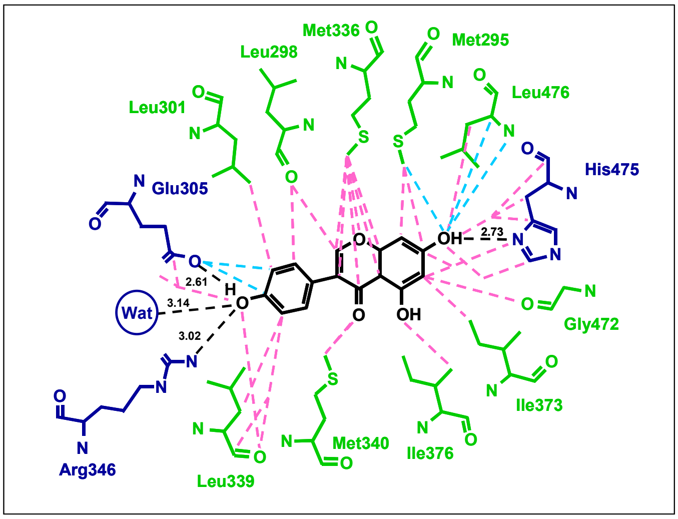
A third amino acid interacts with estrogen's second hydroxyl – but we know interaction while favorable, is not obligatory. So this characteristic only suggests binding affinity – it is not a black and white cut off for a SAR model.



Finally, many interactions involve hydrophobic side chains or the polyamide core – this suggests

a hydrophobic characteristic for ligands and hydrophobicity is necessary to passively diffuse across the target cell membrane. This suggests a certain spatial characteristics for ligands.

However, different ligands have different interactions – the estrogen receptor is relatively flexible and promiscuous. And this leads to a fairly large number of candidates being identified – as a later presentation of the NCTR QSAR model will indicate.



In vitro assays are an unresolved area where further progress is needed:

- several candidate assays exist
- each has its own advantages and limitations
- some are well developed and others are not
- some methods have no common protocol , such as the transfected cell line, the cDNA used with its promoter and its reporter gene

Briefly - Receptor binding is the essential mechanistic step, but will not distinguish agonist from antagonist – or suggest if a compound is a full or partial agonist.

- Yeast reporter gene assays are transcription based, but are not true mimics of the vertebrate steroid transcriptional systems – but are relatively easy to culture and handle
- Vertebrate reporter gene lines are more complex – they may confront transfection stability and sensitivity, but offer realistic transcription responses in some cases.
- Other assays such as MCF-7 proliferation, present issues of specificity, a relatively high rate of false positives even indicating that ethanol is estrogen in an international round robin.

We urgently need comparisons and validation of these *in vitro* assays.

Today, the USEPA has a binding affinity database of over 200 chemicals using a standardized protocol.

- The data base is being expanding with another 250 chemicals, and efforts to validate the protocol are underway.
- There is also a data base on about 500 chemicals here in Japan generated by the high through put screening program.
- Once compared, these data bases would provide the means to validate both QSAR models and other *in vitro* assays.

Animal welfare organizations have sometimes opposed endocrine programs – so after using QSAR and *in vitro* screens – the need for *in vivo* screens must be explicit– metabolism and toxicokinetics are available only in the intact animal – for example.

- Methoxychlor is activated by demethylation in the liver –
- Other compounds may be activated by hydroxylation – or deactivated by esterases –
- Conjugation with glucuronides or sulfates speeds excretion and renders the compound unable to cross the cell membrane and bind the receptor.
- Finally, the hydrophobicity of most estrogen agonists may lead to partitioning.

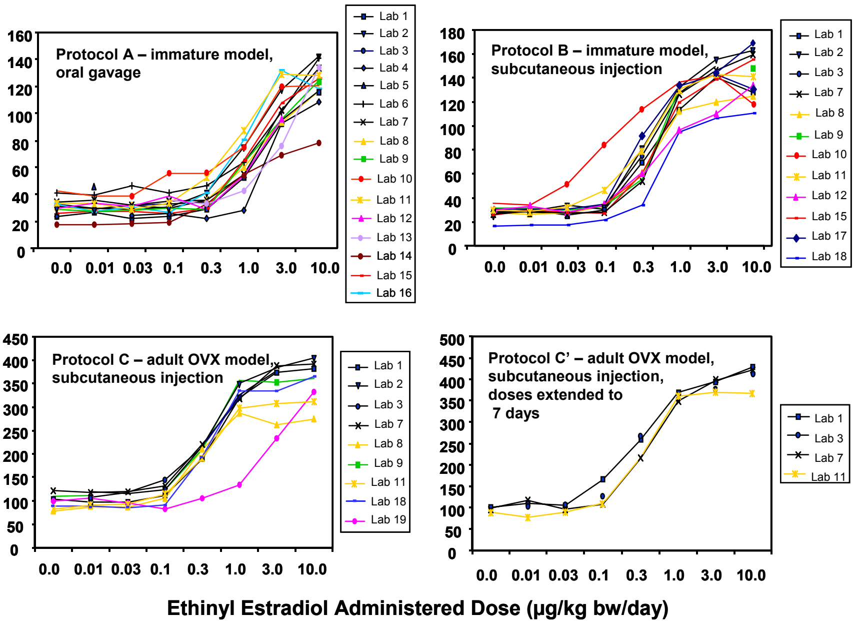
Thus, we have to use animals in *in vivo* screens for relevance before proceeding to large scale tests.

The uterotrophic bioassay is an ideal bioassay:

Estrogen regulates the growth of the target tissue – and in rodents, uterine growth is rapid and dramatic with 4-4.5 day cycle, and an estrogen-stimulated growth phase of ~ 2 days; the dynamic range of the growing uterus is up to a 5-6 fold weight increase, and the measurement is continuous and quantitative. So the basic assay conditions are: 3 days administration, a small group size of n=6 is sufficient, and one needs a control plus 3 doses: so only 24 animals are needed.

First let me acknowledge that the lead laboratory for this OECD activity was from Japan – and express my sincere thanks to Drs Inhoe and Kanno. In phase I of this validation, protocols for the immature and the ovariectomized young adult were standardized. Then these protocols were successfully demonstrated with a potent reference estrogen, ethinyl estradiol. In phase II, these tested protocols were then demonstrated on several weak agonists, the specificity was tested with a negative chemical, and both a dose response and replicability within and among laboratories were demonstrated.

These are the phase I results for blotted uterine weights, at half log increments of EE, among some 19 labs from ten nations. The top half is the immature protocol using oral gavage on the left and sc administration on the right. The bottom half is the ovariectomized young adult after 3 days of sc administration on the left and 7 days on the right.



Given natural biological variation and that some of these labs were performing the uterotrophic bioassay for the first time, I and other members of the validation management committee consider the replication to be excellent and the two basic protocols equivalent.

This slide shows the binding affinities of several compounds, including the EE reference, and several weak agonists used in phase II – the weak agonists represent the likely regulatory targets of the uterotrophic assay and are two to five orders of magnitude lower binding affinity.

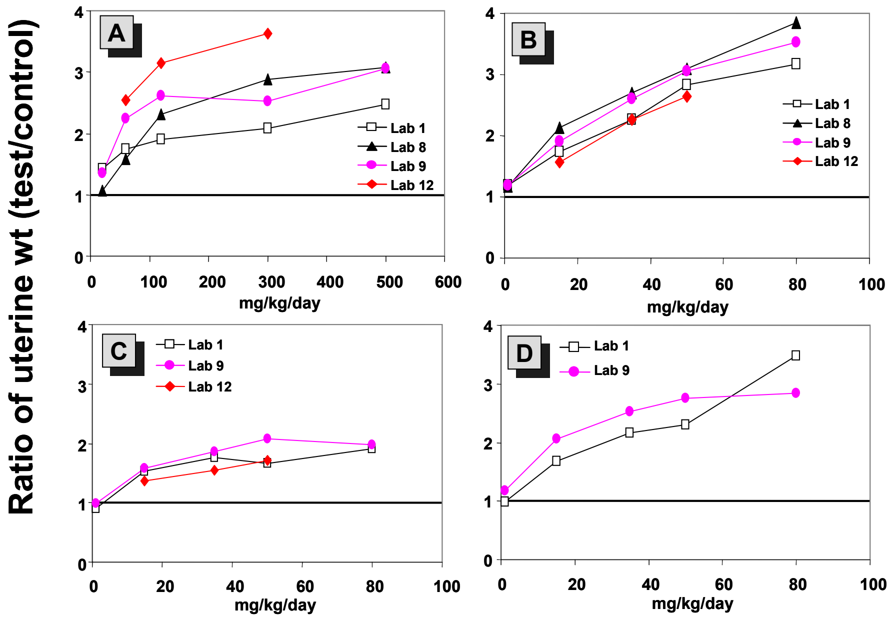
The data include the active metabolite of methoxychlor – HPTE. All data are from the same lab – the NCTR lab where the ER assay has been standardized and QSAR models developed from a database of more than 200 compounds.

Chemical Name	Mean IC_{50} (M) ± S.E.M.	RBA	Log
(Abbreviation)		(%)	RBA
17β-Estradiol (E2)	$8.99 \times 10^{-10} \pm 0.27 \times 10^{-10}$	100.000	2.00
Ethinyl Estradiol (EE)	$4.73 \times 10^{-10} \pm 0.60 \times 10^{-10}$	190.063	2.28
Genistein (GN)	$2.00 \times 10^{-7} \pm 0.21 \times 10^{-7}$	0.443	-0.35
Dihydroxymethoxychlor(HPTE)	$3.55 \times 10^{-7} \pm 0.15 \times 10^{-7}$	0.253	-0.60
Methoxychlor (MX)	$1.44 \times 10^{-4} \pm 0.66 \times 10^{-4}$	0.001	-3.20
4-Nonylphenol (NP)	$3.05 \times 10^{-6} \pm 0.15 \times 10^{-6}$	0.029	-1.53
Bisphenol A (BPA)	$1.17 \times 10^{-5} \pm 0.64 \times 10^{-5}$	0.008	-2.11
o,p'-DDT	$6.43 \times 10^{-5} \pm 0.89 \times 10^{-5}$	0.001	-2.85

These are phase II results for the phytoestrogen genistein. Protocols are in the same position on the slide.

The results are plotted relative to the controls at a starting value of 1, indicated by the horizontal line. Doses are now in tens and hundreds of mgs/kg/day rather than micrograms.

As expected from pharmacokinetic data, greater oral dosage is required as >95% of the orally administered genistein in the serum is conjugated and inactive as these charged forms cannot diffuse across the target cell membrane. Similar results were seen for bisphenol A. In the case of nonylphenol, similar route differences were seen, but were not as dramatic.

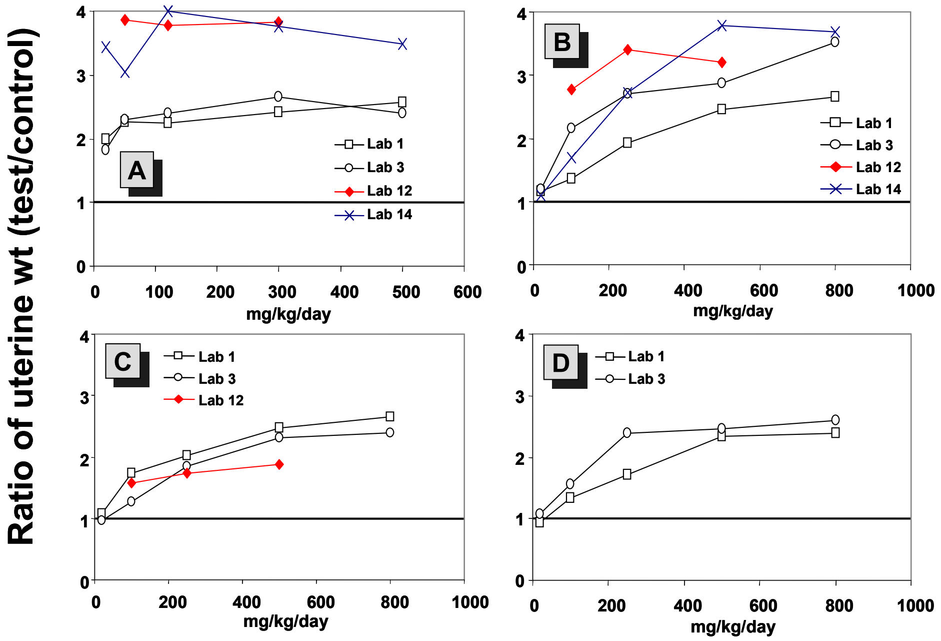


With methoxychlor, oral gavage leads to rapid activation and a lower effective dose versus sc administration. To our mild surprise, *o,p'*-DDT was also more effective by the oral route in all laboratories.

Due to time, I must quickly summarize the other results. The overall replication was excellent – including successful replication in a blind or coded multichemical stage.

With the negative chemical, we did note that 3 labs had slight increases that were statistically significant, but also balance by 2 labs with slight weight decreases that were also statistically significant – indicating some background variation and possibility for false negatives and false positives.

These incidents are being carefully evaluated. Besides replication, an essential question for validation – are the results relevant and predictive?



This summarizes the uterotrophic dose that first achieved statistical significance – the minimal effect dose - and compares it to test data LOELs in the right column. This uses oral gavage data to compare to dietary studies.

As you can see, the uterotrophic assay appears to be a good predictor – when the effect appears to be estrogen related (the compounds in italics) But cannot account for other primary toxicities observed in reproductive and developmental assays (the compounds in bold).

There is also exposure to consider – the apparent margin of safety varies widely – human consumption levels of genistein are low mgs/kg/day, contrasting to micrograms or less for some of the other compounds.

All doses are in mg Substance/kg Body Weight/day

	Uterotrophic MED	LOEL/LOAEL	Effects
	(oral gavage)	(dietary)	
Methoxychlor	*< 20*	*5*	*(vaginal opening)*
Genistein	*~20*	*75*	*(vaginal opening)*
		50	*(latent cervical cancer)*
Nonylphenol	*30-75*	*68*	*(vaginal opening)*
Octylphenol	**> 200**	**300**	**(BW ↓ & organs ↓)**
Bisphenol A	**400-600**	**50**	**(BW ↓)**

Now, an important question –

As the uterotrophic assay covers a single mechanism, estrogen agonists and antagonists, and multiple endocrine mechanisms exist: Must a large and expensive battery be developed to cover all mechanisms?

Or is a comprehensive bioassay necessary – and must it be theoretically complete?

Recognize that such an assay is unlikely to simple and reliable screen – but complex – and its sensitivity is unlikely to be equal to highly specific pharmacological screens such as the uterotrophic and hershberger.

Several more complex subchronic assays or "mini-test" approaches are being evaluated.

These subchronic assays attempt to cover more than one endocrine mechanism with a battery of apical endpoints using tissue weights, histopathology, developmental landmarks such as vaginal opening and so on.

- OECD: Modified 407 28-day Repeat Dose 28 days; 40-80 animals
- USEPA and others: Pubertal assays 40+ days; 120 animals each sex
- USEPA: *in utero* and lactational assay 65 days; 500+ animals
- ACC and others: Intact male assay 15 days; 60 animals

However, the "mini-tests" in using apical endpoints make a mechanistic profile or fingerprint difficult, if not impossible. The lack of specificity may lead to a large number of compounds being labeled as 'Endocrine Disrupters'.

No international harmonization exists here. The USEPA pursues pubertal and *in utero* exposures and the OECD is evaluating the modification and enhancement of the 407 28 day repeat dose assay. Given the animal numbers and expense of these assays, this duplication is worrisome – and the USEPA has described its subchronic assays as screens and not tests. So there is no harmonized framework here.

There is also new and entirely exploratory approach – toxicogenomics.

The concept is fingerprints or profiles based on multiple genes involved in the mechanism of action at the molecular level – not just with 4 or 5 genes, but 60 to 100 genes to describe a given endocrine mechanism.

The challenge– is how to link such up and down regulation of genes to adverse effects for proper interpretation. This means that a dose response and the temporal pattern of gene expression must be understood, and the mechanism deciphered - as one gene product may lead to the transcription of a second or even third wave of genes. Again, very exploratory – but potentially rapid and specific – and a number of mechanisms could be analyzed on a single array chip within a few days and using a limited number of animals.

To state the conclusions of this presentation:

- We have made much progress in the last 3 years, particularly for estrogen screening.
- There are sufficient and sensitive endpoints for estrogen in the definitive 2-gen protocol.
- QSAR models are being developed and can soon be compared and, hopefully, validated.
- The uterotrophic validation is nearly complete and awaits independent peer review.
- The weakest area is the *in vitro* assays. These assays are numerous, in various stages of development – and validation activity is limited.
- However, programs using short-term tests such as the 407 or a pubertal assay have not yet been harmonized.
- Still, the data show that a tiered approach to estrogens is technically feasible – and should reduce the time, expense, and number of animals.
- But international cooperation and harmonization is necessary at all tiers – large amounts of time and resources will be wasted if different methods are used or interpreted differently.
- We need to encourage the many regulatory agencies to find common ground and to harmonize and validate this approach in an international framework. We also need to encourage the sharing of work to move faster and to reduce expense.
- And as noted by Dr Koeter, we have the means to harmonize and validate using the OECD – we have the way to reach a successful conclusion on this effort.

Thank you very much for your kind attention, are there questions?

Q&A

Koëter: Can you work towards your conclusion?

Owens: I beg your pardon?

Koëter: Can you work towards your conclusion?

Owens: Yes. In cases of the predictivity here, methoxychlor, genistein, nonylphenol, this is the minimal effective dose. Here are those from dietary administration, oral gavage. You can see that in many cases there is a very close prediction. However, when the toxicity that first appears, the primary toxicity is not one that is related to the estrogens; the predictivity begins to break down and these tend to be at lower doses, then were found with the uterotrophic.

Let's move ahead. Single mechanisms: we need a mini-test approach possibly, but here, the time and the number of animals begins to increase rapidly.

One of the things that they also do is to use apical end points, but these may lack specificity so that a profile may be difficult. One of the ways of asking in the future is a better way going to merge for doing multiple mechanisms because it is the purpose of the mini-test to do multiple mechanisms, rather than a single one.

They are, however, new and untested, exploratory: you have to establish a multi-gene finger print or profile, and there is the question of how we are going to link these to adverse effects in order to interpret them. There is also the need to do dose response and temporal pattern of gene expression to decipher the mechanism. But again, this would be far more rapid and more efficient and less costly.

The conclusions that Herman asked for, the methods the regulators are needing to address the endocrine issue are becoming available. We are making progress; in fact, this has all come together almost at light speed for some of the other Test Guidelines. We have sufficient and sensitive endpoints in the two-gen; QSARs are in development and are being validated; the OECD program for the uterotrophic is complete.

But the *in vitro* assays are in various stages of development, and we also have the issue of the mini-test, where we do not currently have international harmonization. I would point out that the programs are starting to use common chemicals so that the data can be compared.

So the tiered approach for estrogens appears feasible. It does reduce time, expense, and animals, in theory; however, the previous and excellent international cooperation and harmonization for all tiers need to be continued. This also includes the sharing of work and data, as Herman has proposed, and remember here the downside.

Failure here to adopt a harmonized, tiered approach means that we will collectively waste time, waste resources, and with different results and interpretations that means we are going to have disputes over chemicals, and it can be avoided.

To close out: we are very close here on the estrogen program. More progress is still needed, but within I believe the next two years all will be completed. That is it. Thank you.

Koëter: Thank you very much.

Owens: Now, questions?

Koëter: Everything seems to be very clear this afternoon. Do you have a question? Yes, please.

Q: That was a great talk. You mentioned the QSAR models do not have to be very accurate. Did the OECD ever discuss and provide the guidance about what kind of level of accuracy they are looking for in the QSAR models, particularly with respect to false positive, false negative, and the quantitative predictions?

Owens: The question relates very much in a sense, I believe, to the mixed history that we have for

QSAR models. In a sense, individual parties or institutions have proposed a number of models in the past. When they were then tested, they ran into problems, and so there is a great deal of skepticism about how well QSARs can perform.

I believe that if you take the approach that I am advocating, they need to be carefully constructed to capture prioritized suspect compounds without having to be perfect. That is, allow them to identify compounds that will then need to be assayed either *in vitro* or *in vivo*. You have a more practical assay for them.

In this case, note that a robust set of over 200 compounds has been used for training. I will say in this case that a very broad set of chemical structures and binding affinities as well as negative chemicals, has been used in the training set. And that is key. You cannot go selectively to a few high potency compounds and expect a QSAR model to be robust and be well developed.

Moreover, you notice the second step, which is to go down to a larger data set, of testing chemicals to challenge this with. That provides a way to go to validation. How well it will perform across the range of compounds then becomes a demonstration.

The key aspect of validation is set your success criteria ahead of time, and one of the things that will need to be done is for experts and stakeholders to come together and say this is how we would expect a QSAR to perform. Does it mean it has to predict a binding affinity perfectly? No. Plus or minus an order of magnitude would be a great assistance, and again with a criteria of avoiding false negatives. Under those conditions, that is what I would take forward to a validation program.

Koëter : Thank you very much.