Downstream Users of Chemicals Co-ordination group

DUCC comments to documents CASG-ED/2021/02 and CASG-ED/2021/03
4th Meeting of Competent Authorities Sub-Group on Endocrine Disruptors

---

According to the discussion that took place at the fourth meeting of the CARACAL Sub-Group on Endocrine Disruptors (ED) on March 22nd, 2021 DUCC is submitting the following comments on the documents CASG-ED/2021/02 and CASG-ED/2021/03 on:

1. Inclusion of criteria for endocrine disruption in the CLP Regulation
2. Update of the REACH Annexes in relation to endocrine disruption properties

Introduction

DUCC supports horizontal harmonized ED identification criteria based on the globally accepted WHO definition of Endocrine Disruptors (EDs) across EU legislation. Therefore, it should be fully aligned with the existing criteria already included in the Biocides and Plant Protection Products Regulations.
DUCC disagrees with the approach taken by the Commission to consider CLP as the only option to identify EDs without evaluating other potential options under the REACH regulation. REACH has already demonstrated its ability to identify and assess Endocrine Disrupting chemicals. The implementation of ED identification criteria under REACH could be done in the same way as PBTs via an Annex.

Endocrine activity (EA) is a mode of action that may or may not lead to adverse effects. CLP is designed to communicate hazards and EA is not a hazard per se. Most adverse effects caused via an endocrine mode of action are already captured by existing CLP hazard classes and result in appropriate risk management measures. Therefore, listing EDs under REACH (as currently) would be sufficient to identify ED.

Regarding the update of the REACH Annexes, we think that the two proposals are premature in nature and should be developed further, especially in terms of assays, some of which are partly not yet existing or validated and no criteria for interpretation of results are available. In addition, triggers and waivers which can be derived from existing in vivo studies should be included.

We do not support the obligatory nature of the Annex VII five *in vitro* tests in both Proposals 1 and 2. This requirement exceeds the requirements of the biocides regulation Annex III (where studies on ED "may be included", "but are not limited to…"). Among these are OECD 455 and 458 which cover four endpoints (agonist and antagonist mode). These assays were only validated for reproducibility, not validated for predictivity or relevance, and based on current scientific knowledge would generate a high rate of false positives. This may have serious implications for SMEs only registering at Annex VII, it could lead to drastically increased animal testing, and it brings unresolved issues for the Cosmetic industry as following up on these screening results conflicts with the ban on animal testing for cosmetic products.

It appears pre-mature to implement such a partly validated test battery in REACH information requirements. The **ANNEX** at the end of this document elaborates further on this topic.

1. Draft proposal on hazard classes for endocrine disruptors in CLP

- With the current "**Draft proposal on hazard classes for endocrine disruptors in CLP"** the European Commission (EC) proposes the implementation of new hazard classification & labelling for Endocrine Disruption (ED) for Human Health and the Environment. These new hazard classes are intended to be implemented first in CLP, and then in UN GHS. On the contrary however, implementing the proposed changes in UN GHS and then introducing the hazard categories into the CLP after "global" alignment and "harmonization" appears to be the correct and appropriate process. If the European Commission (EC) however continues to implement them in CLP first, EC should take care that the already internal aligned wording and definitions for ED are used. Therefore, the wording and definitions which are being used under chapters 3.11.1.1-4 in the current draft proposal need to be identical with global and international definitions.

- In general, the wording used in the draft chapter 3.7.2.1.1 has been adopted from the "reproductive toxicity" chapter of the CLP. The problem of double classification for effects of chemicals causing e.g., reproductive toxicity through an endocrine mode-of-action is not clearly addressed here. This problem needs to be addressed in order to avoid over and redundant classification and labelling for one single effect resulting in two separate classifications, probably without differences in the regulatory consequences.

- The current EC proposal for new hazard classes for Endocrine Disruptors follow a CMR-like approach. In table 3.11.1, it is not distinguished between data coming from known evidence in Humans (Cat. 1A) and presumed relevance for humans (Cat. 1B). In contrast to the "classical" CMR endpoints, the proposed Cat. 1 is a "merged" category of Cat. 1A and Cat. 1B. The reason given was that there is no difference in regulatory consequences. However, proven human endocrine disruptors such as Diethylstilbestrol (DES) up to now have rarely been identified. The merged Cat. 1 for EDs appears to ignore the differentiation between human and animal data-based categorization which produces a lack of transparency. This lack of transparency potentially leads to the impression that more evidence (and eventually human evidence) is available for placing a chemical in ED Cat. 1 compared to what evidence exists (for example based only on animal-data). Compared to the "classical" CMR-endpoints, this might lead to a wrong perception if it is not clear whether the evidence for an ED Cat. 1 results from direct human or direct animal evidence with a presumed relevance for humans.

- We support a proposal for an ED Cat. 2 ("suspected") that is scientifically based and requires that an endocrine effect indicated by *in vitro* studies needs to result in an adverse effect *in vivo* and thus in an intact organism. Further emphasis is given to the "plausible link". We think that this point is of utmost importance because a category 2 for "suspected" EDs based on *in vitro* testing results alone would not be scientifically justified and is also in conflict with the hierarchy of endocrine testing methods as laid down in the OECD conceptual framework on ED. Moreover, and the basis for serious concern, the available *in vitro* tests for ED (currently only for "E" and "A", and "S" out of the EATS modalities) have not been designed and used for screening purposes for regulatory classifications so far. The original intention and their current use in the Biocides and Pesticides field is rather a "backwards" - orientated mechanistic research in case of positive findings on ED-parameters reported *in vivo* studies and if there are indications for an endocrine mode of action. For both Biocidal and Pesticidal substances, an extensive (eco-)toxicological dataset is needed which follows the regulatory

principle of an authorization and is not depending on the tonnage-band of a chemical as is the case under REACH (with increasing data-requirement at higher tonnage-bands). So, there is a basic and "regulatory-intrinsic" conflict that needs to be considered by EC in the synchronization.

- We agree with the wording below table 3.11.1 stating that there is also the possibility of non-human relevant adverse effects on the endocrine system. This is important and should be kept within the draft because with more and more testing results and an increasing scientific database, knowledge on the human relevance will increase. Even within the current CLP version, regulatory accepted examples of reproductive toxicity and carcinogenic effects without human relevance and/or reduced relevance due to "secondary" consequence of effects are described. And it is also likely that certain adverse effects on the endocrine system observed in rodents as the experimental animals used will lack relevance to humans. Therefore, the clarification in the chapters 3.11.2.2 and 3.11.2.3 is important to avoid over-classification. Furthermore, paragraph 3.11.2.5 is important in this context acknowledging the fact that inconclusive data and /or species difference should not lead to classification for the CMR-like new CLP-hazard "ED", which is the case for the "classical" CMR-hazards reproductive toxicity and/or carcinogenicity.

- The concentration limits for classification and labelling of identified ED-substances in mixtures should be based on the Reproductive Toxicity endpoint as originally proposed (i.e., 0.3 % for Cat. 1 and 3 % for Cat. 2) in the draft. If scientifically justified and based on *in vivo* data, there is also the possibility besides the Generic Concentration Limits (GCL) to derive Specific Concentration Limits (SCL). It is our understanding, that an SCL can deviate from the GCL resulting in (1) a higher limit (e.g., acknowledging a low potency effect) and (2) a lower limit (e.g., for an adverse effect with higher potency). These deviations can only be derived on a case-by-case basis and should be based on sufficient *in vivo* data. However, the derivation should not follow the assumption that no threshold can be derived for ED-effects, which has no current proven scientific basis. Adopting a non-threshold approach is not in line with the vast majority of current scientific consensus and approaches used by other EU authoritative bodies:
    - An eminent group of leading European toxicologists has recently made it clear that exposure and dose-response should be considered, refuting the no threshold assumption (H. Autrup, et al. J Toxicol Environ Health A 2020, 83, 485, DOI: 10.1080/15287394.2020.1756592).
    - The SCCS, as a European scientific authority, takes a risk-based assessment approach and not a no-threshold approach (e.g., recent SCCS assessment of propyl paraben at: https://ec.europa.eu/health/sites/health/files/scientific_committees/consumer_safety/docs/sccs_o_243.pdf).
    - EFSA, as a European scientific authority, in their assessment of low dose effects of BPA and their risk assessment on BPA, also followed an exposure-based assessment (https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2015.3978) and did not follow a no threshold assumption nor did they confirm the non-monotonic dose-response (NMDR) hypothesis for BPA which largely led to the assumption for non-threshold effects for the ED endpoint: https://www.efsa.europa.eu/sites/default/files/consultation/consultation/draft-opinion-NMDR.pdf

- All pharmacological understanding on receptor-mediated effects clearly indicates that for any receptor mediated effect, thresholds do exist and depend on binding affinity and potency of a chemical.
- These scientific views of European expert bodies are not considered when following a no-threshold approach in proposed new regulations.

- The "weight of evidence" (WoE) should be given high priority in this context. Endocrine regulation and dysregulation can potentially result as a secondary effect from stress and more general toxicity, leading to severe changes in body weight (e.g., such effects have been reported in studies using feed restriction). Additionally, general systemic toxicity and / or alterations in the xenobiotic metabolism are also capable of influencing endocrine homeostasis. A clear differentiation of primary effects on the endocrine system from those that are secondary mediated by other toxic effects (or e.g., at the very high dose only) is needed. Analogously, the WoE approach described in the reproductive toxicity and in the carcinogenicity chapters of the CLP and GHS should be given high priority.

2. Revised proposals for update of the REACH Annexes in relation to endocrine disruption properties

- The EC communicated that both "**Revised proposals for update of the REACH Annexes in relation to endocrine disruption properties**" will be subject to an impact assessment (IA) in 2021. Even though the outcome of this IA needs to be awaited, we anticipate that proposal 2 will lead to a significant increase in animal testing *in vivo* due to the ED *in vivo* mechanistic level 3 studies required at Annex VIII already and the consequence of a potential for over-interpretation of in vitro results. Therefore, we reiterate the support of proposal 1 as already commented in an earlier round. Due to revisions in the current proposal 1, we would however like to comment on some additional points of our previously clear preference of proposal 1.

- ED *in vitro* testing is required from Annex VII onwards in both proposal 1 and 2 (i.e., > 1 ton manufactured/imported per year). Therefore, as commented under point 1 on the inclusion of the new ED hazard class for human health into the CLP, the use of *in vitro* ED testing faces a shift in paradigm: from mechanistic verification of effects observed *in vivo* towards screening of chemicals.

- REACH chemicals in higher tonnage bands (Annex VIII to Annex X) already require *in vivo* data as testing requirement. The identified problem is therefore assumed to be significant for the lowest tonnage band chemicals at Annex VII:

  - <u>Annex VIII to Annex X chemicals:</u>
    Some *in vivo* data will be available from OECD 421/422 and/or OECD 407 (or similar studies via the relevant route of exposure) when registering at this tonnage bands. Thus, in case of (a) positive result(s) from an ED *in vitro* study (e.g., on the "E" and / or "A" out of the EATS modalities) at Annex VIII, these *in vivo* studies contribute to the WoE and provide information if an adverse effect on the endocrine system is observed *in vivo* as well to establish a plausible link

(as required by the CLP, please see proposal above). With increasing tonnage band, the WoE and the amount of information increases (up to an OECD 443 "EOGRTS"), allowing a proper assessment of a potential ED effect. In other words, follow-up testing in vivo will be triggered by findings or default requirements in the tonnage band anyhow which will allow an assessment according to WHO definition.

- Annex VII chemicals:
  In the lowest tonnage band and Annex VII, however, the proposed screening using ED *in vitro* tests will cause conflicts with animal testing bans and requirements as no *in vivo* studies are foreseen as a potential follow-up for "positive" in vitro findings. *In vivo* follow-up studies might have an impact on especially lower tonnage chemicals and especially if such chemicals are used (besides their production falling under REACH) for exclusive cosmetic uses.

- Moreover, the available ED -*in vitro* assay has NOT been validated for predictivity and OECD guidelines emphasize that they should not be used for regulatory decision making. The proposed in vitro testing battery will produce a very high number of "false" results with both amplitudes – "false negative" which is concerning for regulators AND "false positive" results, which is concerning as well with respect to a high number of *in vivo* follow-up studies. Therefore, further validation to study the predictivity of these assays is absolutely needed before they can be added to Annex VII. More details and a full analysis are provided in the ANNEX to these comments (see page 6). Otherwise, this proposal may have serious implications for SMEs only registering at Annex VII, it could lead to drastically increased animal testing, and it brings unresolved issues for the Cosmetic industry as following up on these screening results conflicts with the ban on animal testing for cosmetic products.

Brussels, 26 April 2021

About DUCC

DUCC is a joint platform of **11 European associations** whose member companies use chemicals to **formulate mixtures** (as finished or intermediary products) for professional and industrial users, as well as for consumers.

DUCC focuses on the downstream users' needs, rights, duties and specificities under **REACH** and **CLP**.

DUCC's membership represents several important industry sectors, ranging from cosmetics and detergents to aerosols, paints, inks, toners, pressroom chemicals, adhesives and sealants, construction chemicals, fragrances, disinfectants, lubricants and chemical distributors industries. Altogether, their membership comprises more than **9.000 companies** across the respective sectors in Europe, the vast majority being SMEs. The calculated turnover of these companies is more than **215 billion euros** in Europe.

For more information on DUCC: www.ducc.eu
Jan Robinson – DUCC Chair, jan.robinson@aise.eu
Cristina Arregui – DUCC Vice-Chair, carregui@ifrafragrance.org
Lina Dunauskiene – DUCC Platform Manager, lina.dunauskiene@aise.eu

DUCC's public ID number in the **Transparency Register of the European Commission** is: **70941697936**

ANNEX

Tests used outside of their purpose - use of OECD *in vitro* ED tests in ANNEX VII?

Analysis of validation status, data interpretation criteria and positivity rate of ER and AR reporter assays

Highlights

- A battery of five *in vitro* tests is proposed to be added to REACH information requirements, ANNEX VII
- Among these are OECD 455 and 458 which cover four endpoints (agonist and antagonist mode).
- These assays were only validated for reproducibility.
- They were not validated for predictivity or relevance and based on current knowledge would generate a high rate of false positives.
- The data interpretation criteria in these guidelines are not based on a scientific rationale or any public scientific evidence.
- The guidelines are 'performance-based' – but 'performance-based' is only meant in relation to other *in vitro tests* and not vs. *in vivo* relevance.
- Based on an analysis of the Tox21 dataset, the two assays would likely generate > 30% positive calls.
- Combined with other assays in the proposed *in vitro* battery, this rate is likely to further increase and is currently unknown.
- These tests were developed for prioritisation purpose but are proposed here for a different use.
- It is unclear for what purpose these screening data would be generated in "revised proposal 1", and there is no clear rationale why these data should then be generated at Annex VII level.
- However, in "revised proposal 2", it is indicated "*Appropriate in vivo mechanistic studies in Annex VIII must be conducted or may be required by the Agency in case of a positive result in any of the in vitro mechanistic studies*". Thus, in this proposal a large number of animal tests would be triggered by the high positivity rate.
- This may have serious **implications for SME** only registering at Annex VII, it could lead to **drastically increased animal testing**, and it brings unresolved issues for the Cosmetic industry as following up on these screening results **conflicts with the ban on animal testing.**
- **It appears pre-mature to implement such a partly validated test battery in REACH information requirements.**

Introduction

The identification of endocrine disruptors (ED) was always strongly dependent on *in vitro* assays – historically, the term ED was only coined at the Wingspread conference with and as a consequence of the availability of *in vitro* tests to measure estrogen- and androgen-like activity in *in vitro* systems.

Thus *in vitro* reporter assays were also proposed to the OECD test guideline program and are now covered in a number of test guidelines (OECD, 2016; OECD, 2020b). It is often stated that these are *validated* and *performance-based* assays, which are ready to be used e.g., in the REACH regulatory setting – and in the context of the Chemical Strategy, discussions are now ongoing whether such tests should become mandatory information requirements for any new chemical in the EU above a volume of 1 ton (REACH Annex VII).

The validation status of the tests in OECD Test guidelines 455 and 458

To many stakeholders it is not clear what *validated* means in this case: Validation studies are conducted to assess reliability (i.e., the extent of intra- and interlaboratory reproducibility) and relevance (i.e., the ability of the test method to predict or measure the biological effect of interest in an organism) (Hartung et al., 2004).
The estrogen and androgen reporter gene assays covered under OECD 455 and 458 (OECD, 2016; OECD, 2020b) were indeed tested in detail for their intra- and inter-laboratory reproducibility with very favourable results. However, these are only two modules within a full test validation (Hartung et al., 2004). The aspect of relevance and predictivity to the *in vivo* situation is of key importance and has been assessed in detail in the validation of other *in vitro* OECD tests such as the ones introduced for skin irritation and corrosion, eye irritation and skin sensitization (ECVAM, 2014; Spielmann et al., 2007). The resulting test guidelines thus contain a 'prediction model' which separates chemicals into classes, based on the prediction of an apical endpoint (e.g., whether the chemical is predicted to be a skin irritant or not).

On the other hand, the predictivity module **in comparison to *in vivo* data was not a key part of the OECD validation of the reporter gene assays for endocrine activity**. Sensitivity, specificity, and accuracy values were mainly calculated for these *in vitro* tests **when compared to other *in vitro* tests.** Below is a summary of the evaluation for predictivity in the validation reports for the three assays in TG 455 as one example.

- For the BG1Luc Estrogen Receptor, the ICCVAM validation study (ICCVAM, 2011) assessed 42 recommended substances vs. an ICCVAM consensus call. Among these, 35 could be assessed based on the experimental data. However, these 42 chemicals reference calls are mainly **based on *in vitro* data and not from an evaluation of *in vivo* data.** In addition, the validation report contains data from the uterotropic assay on 13 chemicals among these are only 2 negatives (one being correct-negative and one false-positive in the *in vitro* assays, see Tables 5-14 and 5-15 in the validation report, (ICCVAM, 2011)). Therefore, the BG1Luc Estrogen Receptor assay cannot be considered as test validated for predictivity of *in vivo* data based on the validation report.
  In addition, a minimal efficacy of 20% vs. the efficacy of estrogen at a non-cytotoxic concentration not leading to limited solubility up to a maximal test concentration of 1000 µM leads to a positive call. There is actually no background literature indicating that such a weak partial agonistic activity at such high concentration has any *in vivo* relevance. Thus, these decision thresholds in the data interpretation procedures had never been scientifically scrutinized.

- For the ERα CALUX assay, the validation study only compared classification against *in vitro* data from the other validated *in vitro* assays. However the validation report cites an *in vitro* to *in vivo* correlation analysis, which tested 30 synthetic estrogens and other hormone derivatives both *in vitro* and *in vivo* (Sonneveld et al., 2006). This reference indeed reports a very elegant and comprehensive study to estimate potency prediction of synthetic hormone analogues by the *in vivo* method. However, it did not include typical industrial chemicals the assay is intended to be used for, nor did it test classification accuracy based on the data interpretation criteria set down in TG 455. Therefore, also the ERα CALUX assay cannot be considered as a test validated for predictivity of *in vivo* data.

- Finally, the predictivity for uterotrophic data was best studied in the validation report on the hER-HeLa-9903 cell line assay (OECD, 2015), which included an *in vitro* to *in vivo* correlation for 48 chemicals with uterotropic assay data, among them 16 *in vivo* negatives. This analysis yielded a high predicitivity (91% sensitivity and 88% specificity). However, this result was only obtained if chemicals were rated positive in case of at least 50% receptor activation as compared to E2 at a maximal test concentration of 10 µM (Table 16 in the validation report). The final data interpretation criteria included in the test guideline, though, is based on a threshold of 10% activation: With this criterion for efficacy (but still with a maximal test concentration of only 10 µM), the specificity of the assay drops to 50% (calculated from Table 16 in the validation report). Furthermore, as indicated, the validation of the hER-HeLa-9903 assay included a top concentration of 10^{-5} M (The report stated: "*On the basis of sensitivity of the assay system, the concentration range at 10^{-11} -10^{-5}M can detect estrogenic activity of well-known weak estrogenic chemicals*"), however the OECD TG now includes a top concentration of 1000 µM for this assay (unless chemicals are insoluble or cytotoxic at this concentration), i.e. a 100-fold higher concentration than in the test validation. The specificity vs. the uterotrophic assays (which is already low by lowering the threshold to 10% efficacy) will certainly further drop dramatically by increasing the test dose 100-fold. To my knowledge this was not studied nor discussed in any public document and it is unclear why these modifications of the data interpretation criteria of the assay were introduced into the test guideline, other than to predict the outcome of the other *in vitro* assays. It appears that there are no scientific data to justify these changes and no scientific indication that a chemical with 10% efficacy at 1000 µM has any *in vivo* activity. Therefore, also for the hER-HeLa-9903 with the data interpretation procedures implemented in OECD 455, there is no validation for predictivity.

This lack of validation for predictivity was obviously known at the time of writing the test guideline: The OECD guideline thus did not define these thresholds to rate chemicals as positive with the classical term 'prediction model' but used the term "data interpretation criteria". The OECD guidelines 455 and 458 clearly state that "*the outcome of the tests cannot be used on their own for safety assessment decisions and only be used for screening and prioritisation purposes*".

The other term, next to "*validated",* which is also used with different meanings in an OECD context is "*performance-based*". Thus TG 455 is called a performance-based test guideline. The same term was

recently used when validating defined approaches (DA) for skin sensitization (OECD, 2020a). In the case of the DA project, performance-based indeed refers to the evaluation of the DA vs. animal and human reference data, and performance is directly assessed vs. these reference data. For TG 455 on the other hand, performance again refers only to the performance with which a new *in vitro* assay can predict the *in vitro* results of for a number of chemicals tested before in existing *in vitro* assays. Of course, here the cat bites its own tail. This is clearly stated in the relevant documents, but the fact that the term "perfomance-based" is used for predictivity at completely different levels can instil a false trust in what the "*validated, performance-based"* tests in TG455 can now be used for.

Here we only reviewed the status for the estrogen agonist assays, which is the poster child of *in vitro* endocrine assays. It is even less clear what the *in vivo* relevance is for the outcome of the antagonist assays conducted under the conditions laid down in TG 455 and TG458 but conducting and reporting these assays is part of the guideline and would become mandatory with the proposed data information requirements.

Positivity rate of *in vitro* testing: The Tox21 database as a large case-study

What is clear from looking at the low efficacy thresholds for positivity and the high required maximal test concentration (which would only be limited by 1000 µM maximal concentration or cytotoxicity / insolubility) is that a high number of false-positive screening results might be generated, if the OECD tests with their current data interpretation procedures will be implemented as Information Requirements in Annex VII. Currently, we are not aware of reports on the positivity rate of screening random chemicals with the thresholds set in TG455 and 458. However, we can investigate the positivity rate in the Tox21 screening which used one of the cell lines of TG455 and which screened ER and AR agonist/antagonists for 8311 chemicals.
It is important to note, the Tox21 screening routinely tested chemicals up to 80 µM, i.e. at a 12 times lower concentration as compared to the maximal concentration in TG455 and TG 458[1].

Analysis of results from the ER luc BG1 estrogen assay

The Tox 21 screening includes two tests for estrogen receptor agonism/antagonism. As the ER luc BG1 agonist/antagonist assay is based on the same cell line used in TG 455 and hence equivalent, here the results of this assay are evaluated.
Out of 8306 different CAS numbers tested, 937 (11.2%) chemicals are labelled active agonists in the Tox21 ER screening, while 738 (8.9%) are active antagonists. In total 1623 chemicals (19.5%) are either agonist or antagonists. It is not known, how far this value increases by raising the test concentration 12-fold as required in TG455 and for how many chemicals this is possible based on solubility and cytotoxicity, but in general as the test concentration increases, unspecific reactions will certainly increase.
When including the chemicals which are considered inconclusive agonists or inconclusive antagonists, the frequency of positives raises to 17.4% for agonists, 12.6% for antagonists and 28.6% for combined agonist/antagonists. Chemicals are rated inconclusive due to issues with the curve shape or reproducibility, but all chemicals with an "inconclusive agonist" or "inconclusive antagonist" call counted here have either 20% efficacy in agonist or 30% inhibition in antagonist mode, i.e. at least fulfil this decision thresholds of the guidelines, albeit tested only at lower maximal concentration. It is

[1] For the CALUX assays in TG455 and 458 the maximum test concentration in absence of cytotoxicity or solubility issues is 100 µM, while for all other assays it is 1000 µM. It is unclear what the scientific rationale is behind such a difference for assays on the same endpoint.

not clear, how many of these would be rated as positive when applying the criteria in TG455 for curve evaluation, reproducibility, and maximal test concentration.

Analysis of the results of the Tox21-ar-bla assay

The Tox 21 screening includes two tests for androgen receptor agonism/antagonism. The AR mda kbluc assay technically is most closely related to the tests in the TG458 (a classical nuclear receptor reporter gene assay). However, it contains the MMTV promoter containing response elements for both androgen receptor (AR) and glucocorticoid receptor (GR) and may thus be less specific. Some assays in TG458 specifically were designed to limit glucocorticoid response cross-talk. Thus, we here analysed the data for the more specific Tox21-ar-bla assay.

Out of 8306 different CAS numbers tested, 426 (5.1%) chemicals are labelled active agonists in the AR screening, while 1383 (16.5%) are active antagonists, in total 1580 chemicals (19.0%) are either agonist or antagonists. It is not known, how far this value increases by raising the test concentration 12-fold as required in TG458.

When including the chemicals which are considered inconclusive agonists or inconclusive antagonists, the frequency of positives raises to 7.8% for agonists, 20.6% for antagonists and 23.6% for combined agonist/antagonists. Chemicals are rated inconclusive due to issues with the curve shape or reproducibility, but all chemicals with an "inconclusive agonist" or "inconclusive antagonist" call counted here have either 20% efficacy in agonist or 30% inhibition in antagonist mode. It is not clear, how many of these would be rated as positive when applying the criteria in TG458 for curve evaluation, maximal test concentration and reproducibility.

Impact on positivity rate by combining the results of the multiple assays.

The proposed information requirements for Annex VII includes multiple tests. Thus, if we combine all calls for agonists from either the ER or the AR assay, 1206 chemicals (14.5%) are agonists, while 2380 chemicals (28.6%) are labelled as either agonist or antagonist in one of the two assays. This value is raised to 34.6% of chemicals including the inconclusive agonists/antagonist calls. These values on overall fraction of chemicals with a positive rating in any *in vitro* ED screening assay of the proposed test battery almost certainly will further increase because:

(i) the required test concentrations are higher in TG455 and TG458 as compared to the Tox 21 screening and

(ii) since the proposed information requirements ask for further testing on H295R steroidogenesis assay, aromatase inhibition and one or multiple yet to be defined thyroid assays. This extended battery will certainly further raise the overall positivity rate.

Table 1. Summary of the analysis of the Tox21 tests most closely related to tests in OECD TG 455 and 458

	Active (n)	**Active (%)**	**Active/ inconclusive (n)[1)]**	**Active/ inconclusive (%)**
ER luc BG 1 agonist	937	11.3%	1448	17.4%
ER luc BG 1 antagonist	738	8.9%	1048	12.6%
ER luc BG 1 agonist and/or antagonist	1623	19.5%	2375	28.6%
ar-bla-agonist	426	5.1%	647	7.8%

ar-bla-antagonist	1383	16.6%	1713	20.6%
ar-bla agonist and/or antagonist	1580	19.0%	1961	23.6%
ER luc BG 1 agonist / ar-bla agonist combined	1206	14.5%	2271	27.3%
ER luc BG 1 / ar-bla agonist and antagonist combined	2380	28.6%	2877	34.6%

[1)] inconclusives were only counted if they have an efficacy of at least 20% or inhibition of 30% for antagonists in accordance with OECD TG data interpretation criteria

Screening for prioritisation?

The counter-argument to above raised concerns on (i) lack of validation for predictivity, the (ii) lack of scientific validity of the data interpretation procedures /thresholds selected and (iii) the issue of very high positivity rate from the proposed test battery could be that these tests are intended "only for screening and ***prioritisation purposes",*** which actually is the sole purpose of these tests as stated in the OECD guidelines. However, there is a difficulty to this argument:

- Either these tests should be used in the same REACH Annex, in which higher tier studies would also be mandatory to follow up on potential screening results. In such a case, a negative screening result could indeed be used for a waiver of some higher tier studies, even if the (false)-positivity rate is high.
- However, in the current proposal this is not the case (at least not in proposal 1), as Annex VII would require only *in vitro* testing and chemicals would then remain with this screening results including a high rate of false positives until their use volume would trigger registration in a higher Annex.
- In addition, a fixed waiver for higher tier studies **in case of negative *in vitro* results would then be needed to be implemented within the corresponding Annexes.** Else, the assays do not fulfil the prioritisation purpose for which they were intended. In the current draft this is not implemented.

Based on these considerations, it does not appear that the tests are currently implemented in the draft to be used according to the purpose they are described in the OECD TG – and these assays are thus *not fit for the purpose* for which they are proposed here in Annex VII.

Practical implications

- Based on the data interpretation procedures of the OECD guideline, a large fraction of the chemicals, if not a majority, would be rated 'positive' in at least one assay from the proposed *in vitro* battery.
    - In revised proposal 1, it is unclear what this would mean in a regulatory context, but companies would be faced with a drawback in the market based on these results with a high rate of false positives with an unclear scientific meaning. **This would particularly hit small and middle-sized enterprises (SME) who mainly register products under Annex VII** and whose market volumes do not develop to an Annex VIII registration.

- In revised proposal 2, it is stated "*Appropriate in vivo mechanistic studies in Annex VIII must be conducted or may be required by the Agency in case of a positive result in any of the in vitro mechanistic studies*". Thus, in this proposal **a large number of animal tests** would be triggered by the high positivity rate.

- Companies may be tempted to **proceed to higher tier animal studies at a low production volume** (or certain regulatory agencies may actively start to ask for studies in case of positive screening results as stipulated in revised proposal 2) - this would lead to a **massive increase in unnecessary higher tier animal testing** as the false-positive rate *in vitro* is likely to be very high.

- Companies in the **cosmetic sector** will be in a complex situation, as they should not proceed to animal testing, due to the **ban on animal** testing but then will have these screening results in their dossier without a mean to prove non-relevance.

What would be needed to make TG455 and TG458 applicable in a regulatory setting?

It becomes apparent from analysing the validation reports, the OECD test guidelines and the thresholds for positivity and dose-selection introduced into the guidelines, that only the first part of validation (reproducibility) had been completed, and a validation for predictivity would be needed before these guidelines could become mandatory testing requirements for new chemicals. Completely new data interpretation criteria would be required, and an overhaul of the OECD test guidelines needed.

This could be done, and NICEATM in the US has done significant work in this regard. Thus, test results from multiple ER-agonist and antagonist assays were conducted on the ToxCast chemicals, and just looking at a positive label in any of these tests, a large fraction of chemicals are labelled positive (Due to the multiple tests this fraction is even larger than in the analysis of the Tox21 data conducted above).
However, the results had been used to build a model and to calculate an *in vitro* ER score. This was then ranked against highly curated *in vivo* uterotrophic data to derive a decision threshold (Browne et al., 2015). Many of these ToxCast tests have a high positivity rate when used on their own and with the original data interpretation criteria of the original test. However, if these weak responses (leading to a low score) are not considered biologically relevant, prediction of *in vivo* uterotrophic data became possible with a reasonable specificity (specificity raised from 67% to 89%). The original study by Browne *et al.* had a high complexity as it involved 16 tests to calculate the score, however the model was later simplified to only four tests with similar predictivity (Judson et al., 2017). A similar approach would be needed for OECD approved tests to finally validate them also for predictivity and to update their data interpretation criteria prior to their introduction into any regulatory setting.

APPENDIX

Parameters and data source used to analyse the Tox21 data.

The following files form The Tox21 resources were used:

- ER agonists: tox21-er-luc-bg1-4e2-agonist-p2.aggregrated.txt
- ER antagonists: tox21-er-luc-bg1-4e2-antagonist-p2.aggregrated
- AR agonists: tox21-ar-bla-agonist-p1.aggregrated
- AR antagonists: tox21-ar-bla-antagonist-p1.aggregrated

In each case, the column "Channel_outcome" was used. This column integrates data from multiple repetitions and in the antagonist mode also integrates the data from the cytotoxicity counter screen. This has been described in detail in the supporting information of (Huang et al., 2014), Table S3. The databases were filtered for either "active agonist" or "inactive antagonist", and chemicals with multiple entries were calculated as positive if at least one of the entries was positive removing the duplicate call for all those with several entries (several CAS numbers have multiple entries in Tox21). For inclusion of the inconclusives, "active agonist" and "inconclusive agonist" calls were pooled, and a chemical was again counted if it had at least one of these calls. The total number of entries in Tox21 is over 10'000, but all was calculated based on unique CAS numbers, reducing the database to the 8311 chemicals cited in (Huang et al., 2014).

References

Browne, P., Judson, R. S., Casey, W. M., Kleinstreuer, N. C., Thomas, R. S., 2015. Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. Environmental Science and Technology. 49**,** 8804-8814. 10.1021/acs.est.5b02641

ECVAM, 2014. EURL ECVAM Recommendation on the Direct Peptide Reactivity Assay (DPRA) http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/eurl-ecvam-recommendations/eurl-ecvam-recommendation-on-the-direct-peptide-reactivity-assay-dpra, last accessed 21.5.2014.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A. J., Prieto, P., Sabbioni, E., Scott, L., Worth, A., Zuang, V., 2004. A modular approach to the ECVAM principles on test validity. Altern Lab Anim. 32**,** 467-72. 10.1177/026119290403200503

Huang, R., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., Casey, W., Hsieh, J. H., Shockley, K. R., Ceger, P., Fostel, J., Witt, K. L., Tong, W., Rotroff, D. M., Zhao, T., Shinn, P., Simeonov, A., Dix, D. J., Austin, C. P., Kavlock, R. J., Tice, R. R., Xia, M., 2014. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. Sci Rep. 4**,** 5664. 10.1038/srep05664srep05664 [pii]

ICCVAM, ICCVAM Test Method Evaluation Report The LUMI-CELL® ER (BG1Luc ER TA) Test Method: An In Vitro Assay for Identifying Human Estrogen Receptor Agonist and Antagonist Activity of Chemicals. NIH Publication Number 11-7850, 2011.

Judson, R. S., Houck, K. A., Watt, E. D., Thomas, R. S., 2017. On selecting a minimal set of in vitro assays to reliably determine estrogen agonist activity. Regul Toxicol Pharmacol. 91**,** 39-49. 10.1016/j.yrtph.2017.09.022

OECD, REPORT OF THE INTER-LABORATORY VALIDATION FOR STABLY TRANSFECTED TRANSACTIVATION ASSAY TO DETECT ESTROGENIC AND ANTI-ESTROGENIC ACTIVITY. ENV/JM/MONO(2015)34, 2015.

OECD, 2016. Performance-based test guideline for stably transfected transactivation in vitro assays to detect estrogen receptor agonists and antagonists. OECD series on testing and assessment. 455.

OECD, 2020a. Draft OECD Guideline Defined Approaches for Skin Sensitisation. http://www.oecd.org/chemicalsafety/testing/draft-guideline-for-defined-approaches-for-skin-sensitisation.pdf.

OECD, 2020b. Stably Transfected Human Androgen Receptor Transcriptional Activation Assay for Detection of Androgenic Agonist and Antagonist Activity of Chemicals. OECD series on testing and assessment. 458.

Sonneveld, E., Riteco, J. A. C., Jansen, H. J., Pieterse, B., Brouwer, A., Schoonen, W. G., van der Burg, B., 2006. Comparison of In Vitro and In Vivo Screening Models for Androgenic and Estrogenic Activities. Toxicological Sciences. 89**,** 173-187. 10.1093/toxsci/kfj009

Spielmann, H., Hoffmann, S., Liebsch, M., Botham, P., Fentem, J. H., Eskes, C., Roguet, R., Cotovio, J., Cole, T., Worth, A., Heylings, J., Jones, P., Robles, C., Kandarova, H., Gamer, A., Remmele, M., Curren, R., Raabe, H., Cockshott, A., Gerner, I., Zuang, V., 2007. The ECVAM international validation study on in vitro tests for acute skin irritation: report on the validity of the EPISKIN and EpiDerm assays and on the Skin Integrity Function Test. Altern Lab Anim. 35**,** 559-601.