Regulatory Toxicology and Pharmacology 143 (2023) 105461

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Evaluating H295R steroidogenesis assay data for robust interpretation

H. Tinwell [a,*], A. Karmaus [b], V. Gaskell [c], C. Gomes [d], C. Grant [e], T. Holmes [f], A. Jonas [g], S. Kellum [h], K. Krüger [i], L. Malley [j], S. Melching-Kollmuss [k], O. Mercier [g], H. Pandya [l], T. Placke [m], R. Settivari [h], B. De Waen [n]

[a] Bayer SAS, 16 Rue Jean-Marie Leclair, 69009, Lyon, France
[b] Inotiv, 601 Keystone Park Drive, Morrisville, NC, 27560, United States
[c] Nufarm UK Ltd, Wyke Lane, Bradford, BD12 9EJ, UK
[d] BASF SE, Experimental Toxicology and Ecology, Carl-Bosch-Strasse 38, 67056, Ludwigshafen, Germany
[e] Regulatory Science Associates, Kip Marina, Inverkip, Renfrewshire, PA16 OAS, UK
[f] ADAMA Deutschland GmbH, Edmund-Rumpler-Str. 651149, Koeln (Cologne), Germany
[g] Sumitomo Chemical Agro Europe, Parc D'Affaires de Crécy, 10A Rue de La Voie Lactée, 69370, Saint Didier Au Mont D'Or, France
[h] Corteva Agriscience, Haskell R&D Center, 1090 Elkton Rd, Bldg 320, Newark, DE, 19711, USA
[i] HELM AG, Nordkanalstrasse 28, 20097, Hamburg, Germany
[j] FMC, Stine Research Center, 1090 Elkton Road, Newark, DE, 19711, USA
[k] BASF SE, Speyer Str. 2, 67117, Limburgerhof, Germany
[l] UPL Limited, Mumbai, 400051, India
[m] Syngenta, Rosentalstrasse 67, CH-4058 Basel, Switzerland
[n] ISK, De Kleetlaan 12b, 1831, Machelen, Belgium

ARTICLE INFO

Handling Editor: Martin Van den berg

Keywords:
OECD TG456
Steroidogenesis
Endocrine disruption
In vitro
Assay performance

ABSTRACT

The *in vitro* H295R steroidogenesis assay (OECD TG 456) is used to determine a chemical's potential to interfere with steroid hormone synthesis/metabolism. As positive outcomes in this assay can trigger significant higher tiered testing, we compiled a stakeholder database of reference and test item H295R data to characterize assay outcomes. Information concerning whether a Level 5 reproductive toxicity study was triggered due to a positive outcome in the H295R assay was also included.

Quality control acceptance criteria were not always achieved, suggesting this assay is challenging to conduct within the guideline specifications. Analysis of test item data demonstrated that pairwise significance testing to controls allowed for overly sensitive statistically significant positive outcomes, which likely contribute to the assay's high positive hit rate. Complementary interpretation criteria (e.g., 1.5-fold change threshold) markedly reduced the rate of equivocal and positive outcomes thus improving identification of robust positive effects in the assay.

Finally, a case study (positive H295R outcome and no endocrine adversity *in vivo*) is presented, which suggests that stricter data interpretation criteria could refine necessary *in vivo* follow-up testing. Overall, the described additional criteria could improve H295R data interpretation and help inform on how to best leverage this assay for regulatory purposes.

1. Introduction

Chemical-elicited endocrine disruption (ED) is an adverse outcome of both physiological and regulatory importance. It is therefore critical to develop testing strategies that can robustly identify chemicals with potential endocrine activity and ultimately endocrine adversity. One aspect of this testing is to identify chemicals that may interact with steroid hormone production, as steroid hormones regulate biological processes including development and reproduction. The H295R steroidogenesis assay is an *in vitro* system that can inform on endocrine activity by determining a chemical's potential to interfere with the biosynthesis and metabolism of steroid hormones (steroidogenesis). This assay system has been used predominantly to evaluate effects on 17β-estradiol (E2) and testosterone (T) synthesis

* Corresponding author.
E-mail address: helen.tinwell@bayer.com (H. Tinwell).

https://doi.org/10.1016/j.yrtph.2023.105461
Received 10 March 2023; Received in revised form 22 June 2023; Accepted 21 July 2023
Available online 23 July 2023

(OECD TG 456; OECD, 2011a). The H295R cell line is an adrenocortical carcinoma cell line derived from an African American female that maintains the biological characteristics of zonally undifferentiated human fetal adrenal cells (Gazdar et al., 1990; Gracia et al., 2006). It expresses all the key enzymes involved in steroidogenesis and has been used to evaluate effects of chemicals not only on hormone production but also on steroidogenic enzyme activity and expression (Hecker et al., 2006; Hilscherova et al., 2004; Maglich et al., 2014). Recently, a high-throughput adaptation of the H295R assay, measuring hormones and intermediates, was developed as a screening tool as part of the ToxCast program (Karmaus et al., 2016) that can provide insight into chemical effects and prioritize any additional testing if necessary (Haggard et al., 2018, 2019).

The use of the H295R assay in testing for putative endocrine disruptors has been recognized internationally; both the Organization for Economic Cooperation and Development (OECD) and the U.S. Environmental Protection Agency (EPA) Endocrine Disruptor Screening Program (EDSP) have developed test guidelines for using this assay to detect potential chemical perturbation of E2 and T production (Test guideline OCSPP 890.1550 (EPA, 2009a; 2011b; Hecker et al., 2011; OECD Test Guideline 456 (OECD, 2011a)). The H295R steroidogenesis assay is also one of several Level 2 (*in vitro*) assays identified in the OECD Conceptual Framework for Testing and Assessment of Endocrine Disruptors and Guidance Document 150 (OECD, 2018a) as providing information on endocrine mechanisms.

Criteria and guidance for identifying endocrine disruptors have been in force in Europe since November 2018 for all plant protection products (PPP) submissions (new or renewals) (ECHA/EFSA Guidance, Andersson et al., 2018) and European regulatory agencies are now regularly requesting the conduct of Level 2 *in vitro* assays to determine the endocrine activity of PPPs. The guidance reviews the estrogenic, androgenic, steroidogenic, and thyroid modalities, describes a minimum data set to be considered sufficiently investigated, and recommends a stepwise approach to testing. More specifically, this test battery investigates the potential of a compound to interfere with the estrogen and androgen modalities for ED through the conduct of the estrogen and androgen receptor transactivation assays (OECD TG 455 (2021) and 458 (2020) respectively) with the potential to interfere with the steroidogenesis modality being addressed using the aromatase inhibition (OPPTS guideline 890.1200; EPA, 2009b) and the H295R steroidogenesis (OECD TG 456) assays. Follow-up Level 3 *in vivo* assays, consisting of the uterotrophic and Hershberger assays (OECD TG 440 and 441 respectively), are requested if the *in vitro* assays are negative. Furthermore, an *in vivo* Level 5 reproductive toxicity assay is triggered when positive outcomes are observed in any Level 2 or Level 3 assay to be able to contextualize the findings and identify adversity.

With respect to the H295R steroidogenesis assay, concerns have been raised by industry stakeholders about the high incidence of low efficacy, but nonetheless statistically significant, outcomes in the assay and the consequences that such findings could have on further testing requirements. Specifically, this is the only Level 2 ED assay for which there is no short-term *in vivo* (Level 3) assay available as a follow-up. Furthermore, the Level 4 pubertal assays are considered as insufficiently robust to clearly identify adversity due to altered steroidogenesis. Consequently, according to the EFSA/ECHA ED guidance document (Andersson et al., 2018), a Level 5 reproductive toxicity study is required to contextualize any positive signal identified in the *in vitro* H295R assay, which would lead to the use of up to ~1800 animals per *in vivo* study for each test substance concluded to be positive in the H295R assay.

The OECD validation report of the H295R assay recommended using a 1.5-fold change as a cut-off threshold for activity (Hecker et al., 2011); however, this threshold for activity was not included in the original final test guideline (OECD TG 456, 2011a). Thus, based on the discrepancies between the validation report and the test guideline, as well as the regulatory consequences of the H295R assay outcomes that could lead to significant animal testing, a database was compiled by stakeholders comprising of ~16,800 H295R assay data points to explore the nature of the positive assay outcomes. This compendium of data is the first of its kind for the H295R steroidogenesis assay as it allowed a comprehensive evaluation of outcomes using data generated in assays conducted per the test guideline (OECD TG 456). Additional information was also provided for several of the test items included in the database regarding whether a Level 5 reproductive toxicity study was triggered because of a positive outcome in the H295R assay.

During the course of these analyses, the concerns raised by industry stakeholders were echoed by members of the OECD validation management group for non-animal testing (VMG-NA) originally tasked with the validation of the H295R test guideline (OECD 456), which were triggered by the reporting of an unusually high occurrence of positive outcomes by some contract research organisations (CROs) and other laboratories (Jacobs, 2021). Based on their analyses of 20 H295R studies performed at one CRO, the test guideline was recently refined to re-introduce the 1.5-fold change threshold for activity (OECD TG 456, 2011a).

The analyses conducted herein complement this refined activity threshold criteria as they summarize instances of equivocal outcomes with low efficacy (ie $<$ 1.5-fold changes) and provide additional insight into approaches that may complement current test guideline data evaluation interpretation criteria to robustly identify positives that merit *in vivo* follow-up.

2. Materials and methods

2.1. Data sources

Test item data obtained from GLP or GLP-like studies conducted in 7 test facilities and according to the OECD TG 456 protocol were provided by 11 participating agrochemical companies. Raw data were submitted using a standardized template for all test item runs and accompanying quality control (QC) plates. In addition to all control chemical and test item data, the conducting laboratory identity, cell passage number, raw hormone levels for T and E2, and cytotoxicity data were requested. The final inventory comprised 16,841 entries, of which 9975 represent test item plates, which amounted to 50 test items, and 6866 represent QC plates. For the analyses conducted, fold change values and all statistical outputs were computed *de novo* to ensure consistent processing of all data. All analyses were conducted using R v4.1.0 (R Core Team, 2021).

2.2. Run interpretation criteria

The overall dataset comprised 50 test items with 38 of these substances having data from 2 to 4 runs. The analyses described herein focussed on these 38 test items as the data were generated in-line with OECD 456. The remaining 12 test items were not included in the evaluations as only single runs were performed for them. All runs were analyzed and a run interpretation of equivocal, positive, or negative was determined based on fold changes using four different approaches as described below.

The first approach ("Approach 1") was analysis of pairwise concentration-specific significance compared to plate-matched vehicle controls. This was evaluated by ANOVA followed by Dunnett's test as described in the test guideline. In brief, first a Shapiro-Wilks test was used to assess data normality, if data were normal then ANOVA with Dunnett's was run; when data were not normal, Dunn's test was conducted as a non-parametric alternative. Differences were considered significant at $p \leq 0.05$. For the sake of simplicity, results from this workflow are simply called "Dunnett's" in subsequent descriptions.

Additional approaches were integrated into the interpretation from this first "Dunnett's" approach to assess the impact of additional criteria on the individual run and overall final outcome calls. More specifically, Approach 2 integrated the outcome from Approach 1 (Dunnett's) and

used a non-parametric Jonckheere-Terpstra trend statistic to confirm significant concentration-response trend (for either increase or decrease hormone levels) across testing concentrations. Approach 3 (as per the current OECD TG456, 2011a) integrated the outcome from Approach 1 (Dunnett's) with the requirement to achieve at least 1.5-fold change (induction or inhibition), with the mean fold change across replicates per testing concentration being evaluated. Finally, the fourth approach integrated all three of the evaluated parameters in combination: outcome from Approach 1 (Dunnett's), trend test results, and fold change cutoff. A brief description of interpretation criteria for each approach are listed below.

1 Approach 1: "Dunnett's" (ANOVA with Dunnett's test or Dunn's test), per the original OECD TG 456 (2011a)
   a. Positive: two consecutive concentrations significant ($p \leq 0.05$)
   b. Equivocal: one or non-consecutive significant concentrations significant ($p \leq 0.05$)
   c Negative:no significant effect vs. solvent control ($p > 0.05$) at any testing concentration
2 Approach 2: Dunnett's and Jonckheere-Terpstra (JT) trend test
   a. Positive: positive Dunnett's (see criteria 1a above) and significant JT ($p \leq 0.05$)
   b. Equivocal: equivocal Dunnett's and significant JT ($p \leq 0.05$)
   c. Equivocal: positive Dunnett's and insignificant JT ($p > 0.05$)
   d. Negative: equivocal Dunnett's and insignificant JT ($p > 0.05$)
   e. Negative: no significance by Dunnett's and significant JT ($p \leq 0.05$)
   f. Negative: no significance by Dunnett's and insignificant JT ($p > 0.05$)
3. Approach 3: Dunnett's and fold change cutoff, as per the current OECD TG 456 (2022)
   a. Positive: ≥two adjacent concentrations with positive Dunnett's ($p \leq 0.05$) and maximum fold change ≥1.5
   b. Equivocal: equivocal Dunnett's (as described above for Approach 1; point b and c) and max fold change ≥1.5 in at least one testing concentration
   c. Negative: negative Dunnett's even if max fold change ≥1.5 in at least one testing concentration
   d. Negative: anything with max fold change <1.5 for all testing concentrations
4 Approach 4: Dunnett's, JT trend test, and fold change cutoff
   a. Positive: positive Dunnett's, significant JT trend test, and max fold change ≥1.5
   b. Equivocal: positive Dunnett's, insignificant JT trend test, and max fold change ≥1.5
   c. Equivocal: equivocal Dunnett's, significant JT trend test, and max fold change ≥1.5
   d. Negative: anything with max fold change <1.5
   e. Negative: negative Dunnett's, significant JT trend test, and max fold change ≥1.5
   f. Negative: negative Dunnett's, insignificant JT trend test, and max fold change ≥1.5
   g. Negative: equivocal Dunnett's, insignificant JT trend test, and max fold change ≥1.5

Individual runs for all 50 test items were first evaluated using the above four approaches followed by approach-specific interpretation combining the interpretation of individual runs to give an overall conclusion, determined based on the decision matrices for possible outcomes defined in the relevant OECD TG 456 (ie decision matrix described in the 2011 version of the test guideline was used for Approach 1 whereas the matrix described in the 2022 version was used for Approaches 2–4). As the decision matrices require at least two independent runs for interpretation the outcomes were finally determined for a total of 38 chemicals where data from ≥2 runs for both T and E2 were available.

2.3. H295R outcome and triggering of new level 5 reproductive toxicity studies

For 36 of the 38 test items additional information was supplied by the participating agrochemical companies, which allowed the consequences of the H295R outcome to be assessed in terms of additional *in vivo* testing.

2.4. H295R outcome compared with in vivo data: A case study

In vivo data for a test item (ID 20) that tested positive in the H295R assay according to the original guideline (OECD TG456, 2011a) were reviewed to identify any consistent findings or any pattern indicating adversity relating to EAS-modalities in the absence of systemic toxicity.

3. Results

3.1. QC plate performance

The OECD TG 456 protocol includes the evaluation of a quality control (QC) plate as part of each test run in order to monitor assay conditions when testing chemicals. In addition to basal production minimums, the performance criteria for the QC plate require that 10 μM forskolin achieve induction of steroidogenesis as measured by ≥ 1.5-fold induction of T levels compared to solvent control and ≥7.5-fold induction of E2 levels. Additionally, 1 μM prochloraz should achieve inhibition of hormone levels relative to solvent control of at least 0.5-fold (e.g., 2-fold decrease) for both T and E2.

The QC plate data accompanying all testing runs were obtained and compiled amounting to ~1600 data points each for forskolin- and for prochloraz-mediated effects. The log2 fold change of T and E2 levels in response to forskolin and prochloraz were evaluated (Fig. 1). The QC plate data were grouped by test facility for this evaluation (i.e., laboratory where testing was conducted whether on-site or at a contract laboratory) regardless of the data source for this compilation. The compiled data reveal that achieving minimal inhibition/induction with control chemicals on the QC plates is not robustly accomplished for all QC plate data, with several samples below cutoff values for all performance criteria. Nonetheless, the minimal induction of hormone levels by forskolin was generally achieved in most of the data compiled. The inhibition of T and E2 levels by prochloraz on the other hand were less frequently achieved in multiple testing laboratories.

3.2. Evaluation of data interpretation outcomes

After nearly a decade's experience of performing the OECD TG 456 assay, concerns regarding assay outcomes suggest that consideration of alternative approaches for data interpretation may be warranted. The evaluation of multiple data interpretation criteria was undertaken to determine the impact of efficacy cutoffs and of requiring trend testing for concentration-response. First, all test item runs were analyzed *de novo* from raw hormone concentration readouts including determination of relative fold change and statistical evaluation. After integration of cytotoxicity data and conduct of statistical analyses, each run was assessed by four different approaches, one of which was the original OECD TG 456 approach (Approach 1) and a second one used the 1.5-fold threshold cut-off approach (Approach 3), now included in the refined test guideline, to derive an outcome interpretation of positive, negative, or equivocal (Table 1).

The first approach evaluated statistical significance of each testing concentration as per the original OECD TG456 (2011a) defined criteria and yielded the highest number of positive interpretations for the individual runs (66 of 191 total runs evaluated; 36 positive E2 runs and 30 positive T runs). With the integration of either trend testing (Jonckheere-Terpstra trend test; Approach 2) or fold change cutoff (at least 1.5-fold efficacy achieved at one testing concentration, as per the refined

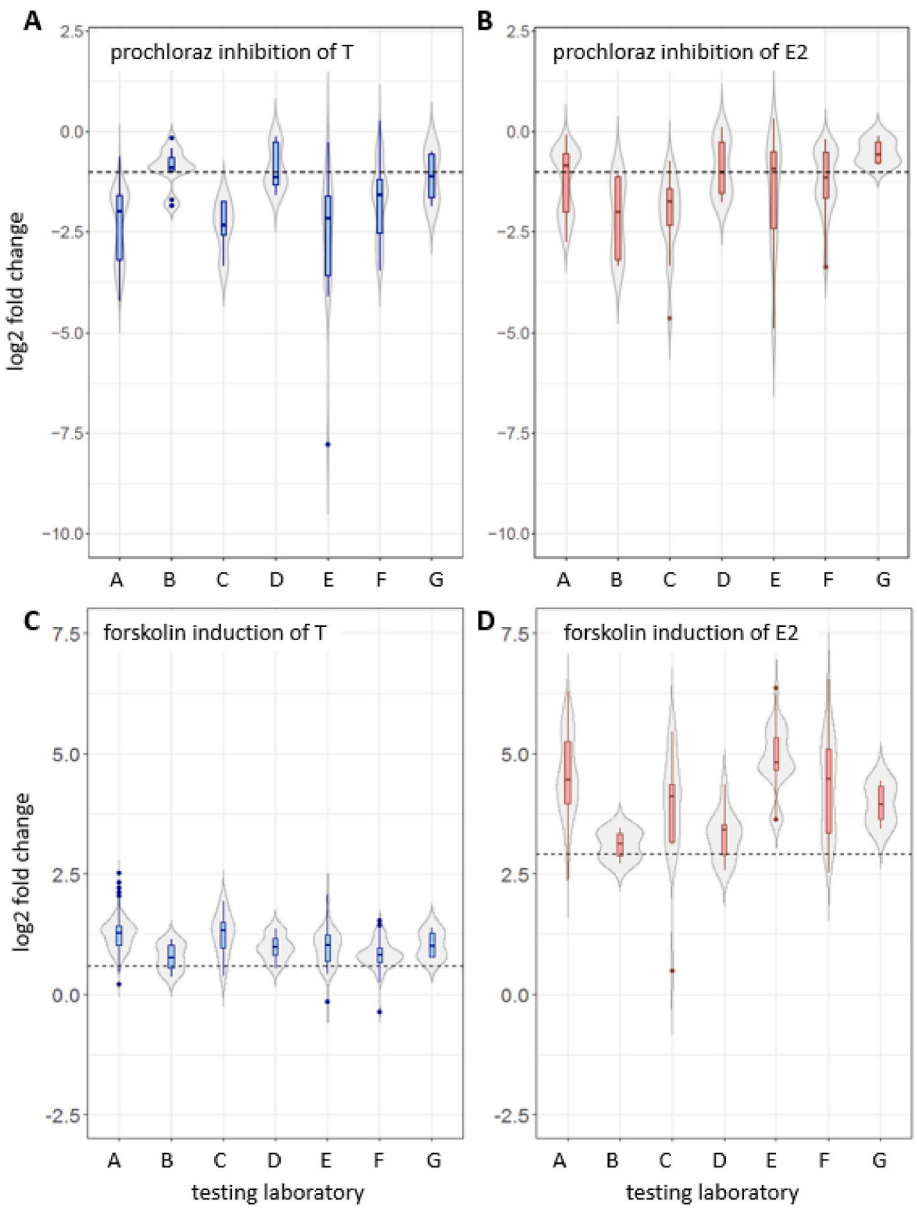


Fig. 1. QC Plate Performance by Test Site. The QC plate performance criteria require evaluation of 1 μM prochloraz-mediated inhibition (**A and B**) and 10 μM forskolin-mediated induction (**C and D**) of T (**blue**) and E2 (**red**) hormone levels as fold change relative to solvent controls. The minimum required inhibition of T and E2 levels is 0.5-fold (in log2 space this is marked as a dashed line at −1 in panels A and B). Forskolin-mediated induction of T should be at least 1.5-fold (marked at 0.58 in the log2 space in panel C) and E2 induction at least 7.5-fold (marked as a dashed line at 2.9 in log2 space in panel D). Data are plotted as box-and-whisker plots to visualize data distribution where the box has a bar representing the median and spans from the 25th to 75th percentile of data points with whiskers extending to the 5th and 95th percentiles, any data points outside this range (likely outliers) are plotted as individual points outside the box-and-whisker plot's reach; a violin plot was added as a shaded area to help visualize data density, with the widest areas revealing where most of the reported data values were.

OECD TG 456 (2022); Approach 3), the number of positive interpretations was reduced. Approach 4, where the statistical significance per testing concentration was combined with both the trend testing and the fold change cutoff did not yield increased stringency above Approach 3 suggesting that fold change cutoff is sufficiently discriminating in identifying robust positives. The most interesting observation, however, was the marked decrease in equivocal calls, and, consequently, the increase in total negative calls as additional criteria were integrated into the interpretation. Notably, the impact of additional criteria inclusion equally impacted interpretations for both T and E2. These results confirm that the addition of trend testing and/or efficacy cutoff to the existing statistical evaluation may help reduce equivocal outcomes and refine the identification of robust positive calls.

Ultimately, test item outcome determinations require evaluating interpretations of at least two independent runs. We derived outcome determinations for the 38 chemicals from our compiled inventory that had at least two runs. Summaries per test item were compiled for all four approaches, for T and E2 separately; results for E2 are provided in Table 2 and for T in Supplementary Table 1. By evaluating the run interpretation and test item final outcomes, it is clearer to see the impact of the additional criteria in using Approaches 2, 3, and 4. Application of the fold change criterion (Approach 3) was more impactful in reducing positive outcomes than simply the addition of a trend test (Approach 2). This is particularly so when considering both overall positive outcomes as well as with the newly defined category of weak positive outcomes (OECD TG456, 2011a).

Detailed results for all runs of an example test item (ID8) that was greatly impacted by the alternative analysis approaches are plotted in Fig. 2. For this test item, using only the ANOVA Dunnett's statistical approach (Approach 1) resulted in runs with equivocal, negative, and positive interpretations, leading, therefore, to the overall conclusion that the test item was positive for interference with steroidogenesis based on the 2011 version of the test guideline. The additional use of the Jonckheere-Terpstra trend testing (Approach 2) showed that two of the three runs had a statistically significant trend for increasing E2 levels; however, in combination with the Dunnett's analysis only one of the runs was considered positive using Approach 2. The overall outcome interpretation for test item ID8 was therefore negative according to the decision matrix of the refined version of the test guideline (2022). Furthermore, all runs were negative when applying the fold change

Table 1
Summary of run interpretations using four different approaches.

		APPROACH 1[b]: OECD TG 456 DUNNETT'S	APPROACH 2: DUNNETT'S AND TREND TEST	APPROACH 3[c]: DUNNETT'S AND FOLD CHANGE CUTOFF	APPROACH 4: DUNNETT'S, TREND TEST, AND FOLD CHANGE CUTOFF
Individual runs[a]	Positive	**66** (36 E2; 30 T)	**50** (27 E2; 23 T)	**28** (17 E2; 11 T)	**28** (17 E2; 11 T)
	Equivocal	**69** (39 E2; 30 T)	**47** (23 E2; 24 T)	**12** (6 E2; 6 T)	**12** (6 E2; 6 T)
	Negative	**56** (21 E2; 35 T)	**94** (46 E2; 48 T)	**151** (73 E2; 78 T)	**151** (73 E2; 78 T)
Overall call per hormone	Positive	**19E; 15T**	**9E; 4T**	**7E; 2T**	**7E; 2T**
	Weak Positive[d]	**NA**	**10E; 10T**	**3E; 3T**	**3E; 3T**
	Negative	**4E; 10T**	**15E; 15T**	**27E; 29T**	**27E; 29T**
	NA	**15E; 13T**	**4E; 9T**	**1E; 4T**	**1E; 4T**
Overall H295R assay conclusion	Positive	**23**	**11**	**8**	**8**
	Weak Positive[d]	**NA**	**11**	**4**	**4**
	Negative	**3**	**9**	**24**	**24**
	NA	**12**	**7**	**2**	**2**

NA: Not Applicable was applied when the decision matrix of the relevant version of the TG does not account for the combination of results (or an additional run was not available in the current dataset which would be required to conclude on an outcome) or the overall final call (ie weak positive does not exist in the original version of TG 456 (2011)).

[a] : sum of all runs for each test chemical having two or more runs (both E2 and T are summed to obtain totals in bold, with hormone specific sums provided in brackets).

[b] Approach as per original OECD TG 456 (2011).

[c] Approach as per current OECD TG 456 (2022).

[d] : The refined OECD TG 456 (2022) includes a new category of weak positive capturing occurrences of equivocal/negative/(positive or equivocal) or equivocal/equivocal/(equivocal or positive) for each of the three runs.

cutoff (Approach 3) as well as Approach 4 highlighting the only marginal increases (<1.5-fold at all concentrations), despite clear trends and high testing concentration significance. This example highlights how critical biological context is for efficacy interpretation and that statistical analyses can support but alone are not necessarily sufficient for interpretation determinations.

3.3. H295R outcome and triggering of new level 5 reproductive toxicity studies

Outcomes from the H295R assay (positive or inconclusive) can lead to the requirement for the conduct of a Level 5 reproductive toxicity study as defined by the OECD Conceptual Framework (OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters). Of the 38 test items that were subjected to the above analyses, additional information was provided for 36 of them in terms of whether a new Level 5 reproductive toxicity study was triggered due to the H295R outcome (details given in Table 3). Fig. 3 illustrates that a new reproductive toxicity study (OECD TG 443: Extended One Generation Reproductive Toxicity Study or OECD TG 416: 2-Generation Reproductive Toxicity Study) was performed for 13/36 test items because of either a positive or inconclusive conclusion according to the original 2011 OECD TG 456 criteria. As indicated earlier in Table 1 and highlighted in both Fig. 3 and Table 3, the incidence of positive/inconclusive outcomes in the assay for these test items was markedly reduced when applying the fold change cutoff introduced into the refined OECD TG 456 (2022).

Although it would have been interesting and pertinent to provide concrete information concerning the false positive rate in this assay with respect to the outcome in the newly conducted Level 5 reproductive toxicity studies, this was not possible because, at the time of writing, most of the 13 test items were undergoing the renewal process in the EU and the official regulatory conclusions on their ED potential were not yet available. More specifically, out of the 13 test items with additional reproductive toxicity data, only one has recently been concluded as non-ED for human health (HH) by EFSA (official EFSA conclusion not yet available) and the draft assessment reports (DAR) prepared by the rapporteur member state (RMS) for an additional two substances indicates non-ED for HH. Of the remaining 10 test items, a conclusion of non-ED (specifically referring to the estrogen, androgen, and steroidogenesis (EAS) modalities) has been proposed by the applicants for 8 test substances and no information was provided for the remaining two compounds (Fig. 3; Table 3).

For the 11/13 compounds considered non-ED (either officially or otherwise) this was due to either the absence of adversity in the extensive database available for each compound (including the new reproductive toxicity study) or due to an absence of a plausible link between the *in vitro* and *in vivo* observations. Interestingly, 4 of the 13 test items for which a reproductive toxicity study was triggered were still concluded as being positive in the H295R assay even after application of the fold change cut-off. Three of these four compounds were considered by the applicants as non-ED due to either the absence of adversity (2 items) or the absence of a plausible link (1 test item); no additional information was provided for the 4th test item. Notably, of the compounds that did not have a new reproductive toxicity study performed, the majority (17/23 test items) would have been negative according to the new OECD TG 456 guideline which integrates fold change, but the one compound that was withdrawn due to ED remained positive. The absence of adversity, despite a positive, albeit weak, outcome in the original H295R assay is described in more detail below using test item ID 20 as a case study.

3.4. H295R outcomes compared with in vivo data: A case study

The weight of evidence assessment to determine the ED potential within the regulatory context, according to the EFSA/ECHA ED guidance (Andersson et al., 2018), for a test item was provided for review herein. This example demonstrates how critical the interpretation of the H295R steroidogenesis assay is because follow-up on a positive outcome requires substantial *in vivo* testing to be conducted.

Briefly, the review of ToxCast/Tox21 high-throughput screening data had initially suggested that the test item (ID20 in the blinded compiled dataset) was negative in the estrogen receptor pathway model (Judson et al. 2015) and androgen receptor pathway model (Kleinstreuer et al. 2017). As, according to the EFSA/ECHA ED guidance (Andersson et al., 2018), the estrogen bioactivity model is considered sufficiently predictive of interference with the estrogen modality, no further testing for this modality was requested; however, more information concerning the potential of the test item to interfere with the androgen and steroidogenesis modalities for ED was required.

Table 2
Outcome interpretation for estradiol (E2) effects per test item.

ID	Approach 1: OECD TG 456 DUNNETT'S	Approach 2: DUNNETT'S AND TREND TEST	Approach 3: DUNNETT'S AND FOLD CHANGE CUTOFF	Approach 4: DUNNETT'S, TREND TEST, AND FOLD CHANGE CUTOFF
1	**Positive:** equivocal, positive	**NA:** negative, equivocal	**Negative:** negative, negative	**Negative:** negative, negative
2	**Positive:** negative, positive, positive	**Positive:** negative, positive, positive	**Positive:** negative, positive, positive	**Positive:** negative, positive, positive
3	**Positive:** positive, equivocal, negative	**WP:** positive, equivocal, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
4	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive
5	**NA:** positive, equivocal, equivocal	**WP:** equivocal, equivocal, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
6	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive
7	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive
8	**Positive:** equivocal, negative, positive	**Negative:** negative, negative, positive	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
9	**NA:** positive, equivocal, equivocal	**Negative:** positive, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
11	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
12	**NA:** Equivocal, equivocal, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
13	**Positive:** equivocal, positive, positive	**Positive:** negative, positive, positive	**Negative:** negative, negative, positive	**Negative:** negative, negative, positive
14	**NA:** negative, equivocal	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
16	**NA:** equivocal, negative	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
17	**Positive:** positive, equivocal, positive	**WP:** positive, negative, equivocal	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
18	**Positive:** positive, equivocal, positive	**WP:** equivocal, negative, equivocal	**WP:** negative, equivocal, positive	**WP:** negative, equivocal, positive
19	**NA:** positive, equivocal, equivocal	**WP:** positive, equivocal, equivocal	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
20	**Positive:** positive, positive	**Positive:** positive, positive	**NA:** negative, equivocal	**NA:** negative, equivocal
21	**NA:** equivocal, equivocal	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
22	**NA:** negative, positive	**NA:** negative, equivocal	**Negative:** negative, negative	**Negative:** negative, negative
23	**Positive:** equivocal, equivocal, positive	**WP:** equivocal, equivocal, positive	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
24	**NA:** Equivocal, equivocal, negative	**Negative:** Equivocal, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
25	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive
27	**Positive:** equivocal, positive	**NA:** negative, positive	**Negative:** negative, negative	**Negative:** negative, negative
28	**Positive:** equivocal, positive, positive	**WP:** negative, equivocal, positive	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
29	**NA:** equivocal, negative, equivocal	**WP:** equivocal, negative, equivocal	**WP:** equivocal, negative, equivocal	**WP:** equivocal, negative, equivocal
30	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
31	**NA:** equivocal, equivocal	**NA:** equivocal, negative	**NA:** equivocal, equivocal	**NA:** equivocal, equivocal
33	**NA:** equivocal, negative	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
40	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative	**Negative:** negative, negative
41	**Positive:** positive, positive	**Positive:** equivocal, positive	**Positive:** positive, positive	**Positive:** positive, positive
42	**Positive:** equivocal, positive	**WP:** equivocal, positive	**WP:** equivocal, positive	**WP:** equivocal, positive
43	**NA:** negative, equivocal, equivocal	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
44	**Positive:** positive, equivocal, equivocal	**Negative:** equivocal, negative, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
45	**Negative:** negative, positive, negative	**Negative:** negative, positive, negative	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
48	**NA:** equivocal, equivocal, equivocal	**Negative:** negative, negative, equivocal	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
49	**NA:** negative, equivocal, equivocal	**WP:** negative, equivocal, equivocal	**Negative:** negative, negative, negative	**Negative:** negative, negative, negative
50	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive	**Positive:** positive, positive

Note: Outcomes for the final interpretation and individual runs for 38 test items for which data from 2 or more runs were available. Final interpretations for outcome scenarios are given in bold and were determined based on the original OECD TG 456 decision matrix for Approaches 1 and 2, with details defined in the Materials and Methods. Approach 3 final interpretations are based on the recently refined test guideline.
NA: Not Applicable was applied when the current decision matrix does not

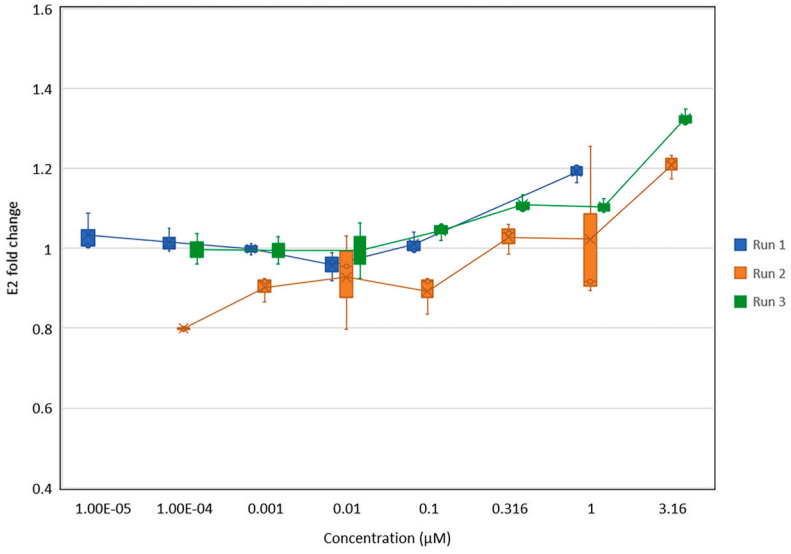
account for the combination of results (or an additional run not available in the current dataset which would be required to conclude on an outcome).
WP: weak positive; the refined OECD TG 456 (2022) includes a new category of weak positive capturing occurrences of equivocal/negative/(positive or equivocal) or equivocal/equivocal/(equivocal or positive) for each of the three runs or equivocal/positive for each of two runs. The same analyses for T are presented in Supplementary Table S1.

Consequently, the AR-EcoScreen™ stably transfected human androgen receptor transcriptional activation assay (ARTA; OECD TG 458; highest concentration tested was 3.16 μM due to precipitation at higher concentrations) and the Level 3 Hershberger assay (OECD TG 441) were conducted to address the androgen modality. The aromatase inhibition (OPPTS 890.1200; highest concentration tested was 100 μM), and the H295R steroidogenesis (OECD TG 456; highest concentration tested was 100 μM based on solubility) assays were performed to address the steroidogenesis modality. No evidence of androgen agonism or antagonism was observed in either the ARTA or the Hershberger assay (conducted at 15, 50, and 200 mg/kg bw/day, based on previous repeat dose studies indicating effects on bodyweight, food consumption, and hematological parameters from 150 mg/kg/d). Furthermore, the aromatase inhibition assay was also negative.

The H295R steroidogenesis assay was the only Level 2 study to elicit any statistically significant positive results. Specifically, statistically significant increased estradiol levels were recorded at two consecutive concentrations in two independent runs in the H295R study (Fig. 4a and b), albeit the maximum fold increases were low (1.44 and 1.56-fold) in both runs at the highest concentration (100 μM). Nevertheless, test item ID20 was concluded to be positive in the steroidogenesis assay according to the original test guideline (OECD TG456, 2011a). Analysis of the data using the additional approaches described herein identified a significant trend test (Approach 2) in both runs thus supporting the statistical analyses of the original TG 456 (2011). Although this test item was concluded as being positive according to the original OECD TG 456 (2011) as well as according to Approach 2 (trend test), application of the criteria in the latest version of the guideline (OECD TG 456, 2022; Approach 3 herein) determined that a third test run would have been necessary to conclude on its potential to interfere with steroidogenesis according to the latest OECD TG 456 (2022) as only the top concentration in a single run was greater than 1.5-fold, thus leading to an equivocal call for this run (Fig. 4b). Furthermore, according to the decision matrix of the latest guideline, the outcome of an eventual third run would have resulted in either a negative or weak positive as the final call (Fig. 4c). Nevertheless, in accordance with the EFSA/ECHA ED guidance document (Andersson et al., 2018) the marginal positive result (as per TG 456 2011) for test item ID20 triggered the conduct of a level 5 extended one-generation reproductive toxicity study (OECD TG 443). The Level 4 female pubertal assay (OPPTS 890.1450 (EPA, 2011a)) was not conducted as the assay is generally considered to be not robust enough to clearly identify adversity resulting from altered steroidogenesis.

To evaluate endocrine disruption potential of test item ID20, relevant endocrine organ endpoints from *in vivo* studies were reviewed and summarized (Table 4). These studies overwhelmingly demonstrate a lack of endocrine effects. The studies included subchronic rat, mouse, and dog studies from which measured weights of endocrine organs and histopathological examinations yielded no evidence of endocrine effects. The main findings were for effects on body weight, with some secondary organ weight changes, effects on food consumption, and on some hematological parameters. Ultimately, the conclusion from these subchronic studies was that there were no effects relevant to estrogen/androgen/steroidogenesis (EAS)-modalities and that the criteria for ED were not met. Additional dietary chronic toxicity and carcinogenicity studies were also conducted, in which no effects on weights or histopathology of endocrine organs were observed, resulting in the conclusion that there was no evidence of EAS-mediated endocrine effects in long-term studies in rats and mice. Finally, data available from several reproductive toxicity studies, including the newly conducted extended one generation reproductive toxicity study, were also assessed for endocrine effects, which despite some parental toxicity at higher doses, showed no consistent effects on reproductive efficiency nor on endocrine organ weight or histopathology.

In summary, the *in vivo* data confirmed that the test item did not provide any consistent findings or any pattern indicating adversity relating to EAS-modalities in the absence of systemic toxicity. The lack of endocrine activity for the test item was confirmed by the absence of effects in the Level 2 (ARTA and aromatase) and Level 3 assays. The additional Level 5 testing initiated based on the marginal efficacy positive call from the H295R steroidogenesis assay in this case did not substantiate the *in vitro* observation (Table 4) and confirmed the lack of



Run analysis details

Run	ANOVA Dunnett's p≤0.05 (μM)	Trend	Maximum Fold Change
1	1	0.39	1.21
2	NA	0.02*	1.26
3	0.316, 1, 3.16	<0.0001*	1.35

Overall outcome interpretation

Run	Approach 1	Approach[A] 2	Approach[A] 3	Approach[A] 4
1	Equivocal	Negative	Negative	Negative
2	Negative	Negative	Negative	Negative
3	Positive	Positive	Negative	Negative
Overall	**Positive**	**Negative**	**Negative**	**Negative**

Fig. 2. Example detailed results for estradiol replicate runs of test item 8 (ID8) in the H295R steroidogenesis assay. The plot shows triplicate observations as boxplots. Tables summarize the statistical analyses and the interpretation derived from each run and analysis approach.
NA: not applicable as no statistical significance observed; *: statistically significant JT trend test ($p \leq 0.05$). A: interpretations for Approaches 2–4 based on the decision matrix in the recently refined TG 456 (2022).

Table 3
H295R assay conclusions and subsequent *In Vivo* testing.

ID	Overall H295R assay conclusion		Level 5 Reproductive toxicity study triggered by H295R Outcome	European ED conclusion for EAS modalities		
	Approach 1: OECD TG 456 (2011)	Approach 3: OECD (2022)		ED	Who	Justification/Comments provided by participating agrochemical companies
1	**Positive**	**Negative**	No	No	Applicant	All modality data negative
2	**Positive**	**Positive**	No	Yes	Applicant	ED activity (withdrawn from Europe)
3	**Positive**	**Negative**	No	No	Applicant	No additional information
4	**Positive**	**Positive**	No	No	Applicant	No additional information
5	**NA**	**Weak Positive**	No	No	Applicant	No additional information
6	**Positive**	**Positive**	Yes (OECD TG 416; 2011b)	No	Applicant	Absence of plausible link
7	**Positive**	**Positive**	Yes[a] (OECD TG 416; 2011b)	No	Applicant	Absence of adversity in higher tier *in vivo* studies
8	**Positive**	**Negative**	Yes[a] (OECD TG 416; 2011b)	No	Applicant	Absence of plausible link
9	**Positive**	**Negative**	Yes (OECD TG 416; 2011b)	No	Applicant	Absence of adversity in higher tier *in vivo* studies
11	**Negative**	**Negative**	No[b]	Unknown		No additional information provided
12	**NA**	**Negative**	No	No	Applicant	Absence of adversity/activity in higher tier studies and Level 3 tests
13	**Positive**	**Negative**	Yes (OECD TG 443)	Unknown		No additional information provided
14	**NA**	**Negative**	Yes (OECD TG 443)	No	Applicant	Absence of adversity/activity in higher tier studies and Level 3 tests
16	**NA**	**Negative**	No (not considered in this analysis as test item was a negative chemical from assay validation study)			
17	**Positive**	**Negative**	No	No	Applicant/ draft ER	Absence of adversity
18	**Positive**	**Positive**	No	No	Applicant	Absence of plausible link
19	**NA**	**Negative**	No	No	Applicant	Absence of adversity
20	**Positive**	**NA**	Yes (OECD TG 443)	No	EFSA (2023)[e]	Absence of adversity (including new repro study)
21	**Positive**	**Weak Positive**	No	No	Applicant	Absence of adversity and activity
22	**NA**	**Negative**	No	No	Applicant	No additional information
23	**Positive**	**Negative**	No	No	Applicant	No additional information
24	**NA**	**Negative**	No	No	Applicant	No additional information
25	**Positive**	**Positive**	Yes (OECD TG 443)	No	Applicant	No additional information
27	**Positive**	**NA**	No	No	Applicant	No additional information
28	**Positive**	**Negative**	No	No	Applicant	Absence of adversity in higher tier studies
29	**NA**	**Weak Positive**	Yes[c]	No	Applicant	Absence of adversity in higher tier studies
30	**NA**	**Negative**	No	Test substance not submitted to Europe		
31	**NA**	**NA**	No	Test substance not submitted to Europe		
33	**NA**	**Negative**	No	Test substance not submitted to Europe		
40	**Negative**	**Negative**	No	Test substance not submitted to Europe		
41	**Positive**	**Positive**	Yes (OECD TG 443)	No	Applicant/ DAR	Absence of adversity (including new repro study)
42	**Positive**	**Weak Positive**	Yes (OECD TG 443)	No	Applicant/ DAR	Absence of adversity (including new repro study)
43	**NA**	**Negative**	Yes (OECD TG 443)[d]	No	Applicant	Absence of adversity (including new repro study)
44	**Positive**	**Negative**	Yes (OECD TG 443)	No	Applicant	Absence of adversity (including new repro study)
45	**Negative**	**Negative**	No	No	Applicant	Absence of adversity
48	**NA**	**Negative**	No	No	Applicant/ DAR	Absence of adversity
49	**Positive**	**Negative**	No	No	Applicant/ DAR	Absence of adversity
50	**Positive**	**Positive**	No information provided			

The H295R assay outcomes based on the original and refined OECD TG 456 criteria (2011, 2022 respectively) are presented alongside the additional information provided by the participating agrochemical companies. A new level 5 reproductive toxicity study (either a 2-generation or an extended one generation reproductive toxicity study) was requested for 13 test items based on the positive/inconclusive outcome in the H295R assay according to the original criteria. The overall conclusion for EAS mediated ED (mainly based on the review of the extensive *in vitro* and *in vivo* database for each compound by the corresponding applicant) are also provided. ER: Evaluation report; DAR: draft assessment report.

[a] Additional Level 2 assays were also concluded as positive.

[b] A positive Level 3 assay triggered an EOGRT (OECD TG 443); however, this study was not considered in the analyses described herein as it was not triggered by the H295R assay.

[c] Study guideline information not provided.

[d] The EOGRT (OECD TG 443) was requested by the US EPA rather than EFSA.

[e] Internal feedback only as official EFSA conclusion not currently available.

ED adversity already concluded from the substantial *in vivo* data package available for the test item. It is interesting to note that this compound was recently concluded as non-ED for HH by EFSA (internal communication; EFSA conclusion not yet published). The inclusion of a fold change cutoff in this case could have, therefore, supported appropriate categorization of this chemical as not having frank effects on steroidogenesis, which would have been consistent with all other findings and mitigated the need for additional *in vivo* studies.

4. Discussion

The H295R steroidogenesis assay was developed as "*a level 2 'in vitro assay, providing mechanistic data' to be used for screening and prioritization purposes*", as stated in the first paragraph of the introduction in OECD TG 456 (2011, 2022). The second paragraph goes on to state, "*the assay does not provide specific mechanistic information concerning the interaction of the test substance with the endocrine system. Research has been conducted using*

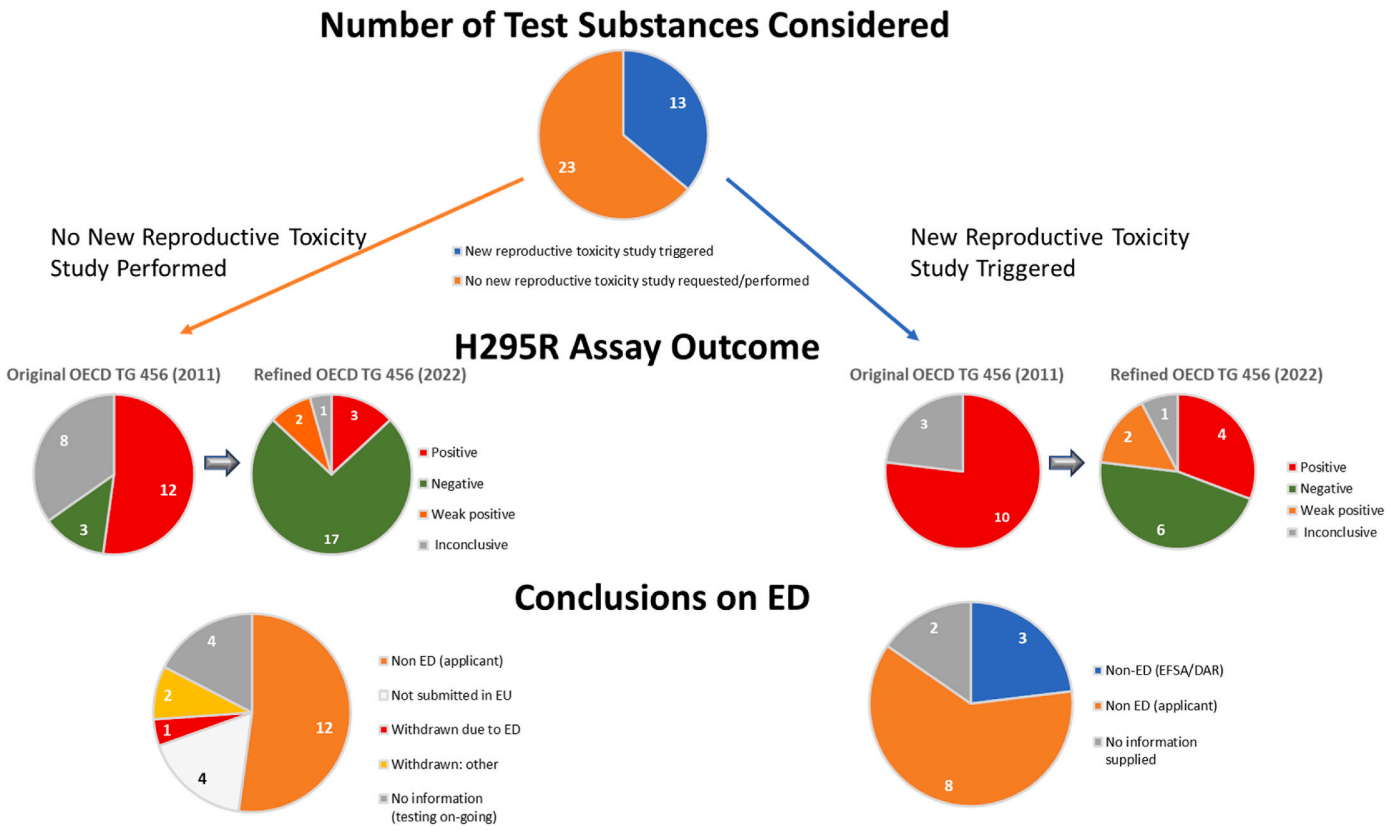


Fig. 3. Additional information was provided by the participating agrochemical companies for 36 of the 38 test items used in the analyses described herein. A new level 5 reproductive toxicity study (either a 2-generation or an extended one generation reproductive toxicity study) was requested for 13/36 test items based on the positive/inconclusive outcome in the H295R assay according to the original criteria. The pie charts provide an overview of the H295R assay conclusions for each scenario (i.e. reproductive toxicity study triggered or not) as well as the overall conclusion for EAS mediated ED (mainly based on the review of the extensive *in vitro* and *in vivo* database for each compound by the corresponding applicant).

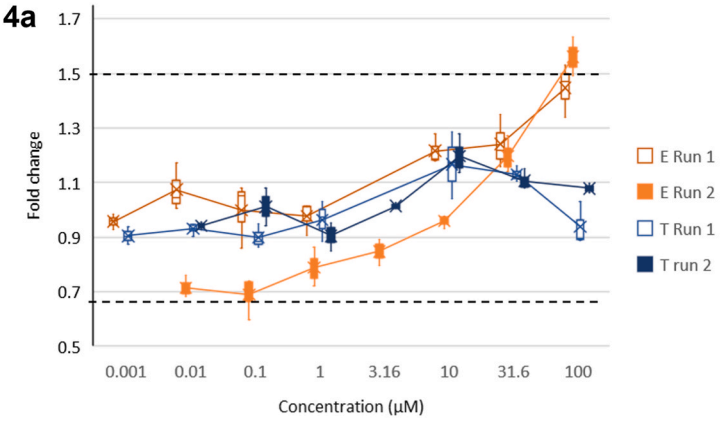
the cell line to identify effects on specific enzymes and intermediate hormones …". This context is critical and was rightfully stated at the forefront of the OECD TG 456 document. Yet, this human cell line is ideal for applications to regulatory testing because it maintains the ability to produce basal hormone levels allowing for the assessment of both induction and inhibition potential of a chemical on steroid hormones. The adoption of this assay for regulatory testing for endocrine disruption potential has led to the results of this *in vitro* assay having a significant impact on animal usage. Positive or equivocal outcomes have triggered substantial *in vivo* testing and concerns have arisen regarding over-interpretation of outcomes from this *in vitro* test system. Thus, the purpose of our analyses was to evaluate the data generated to support regulatory submission of PPPs, to assess outcomes and subsequently provide complementary data interpretation criteria approaches. During the course of our analyses similar concerns raised by CROs and other laboratories led to a re-analysis by the original OECD validation management group of 20 H295R studies performed at one CRO. This re-analysis included use of the 1.5-fold change threshold for activity originally recommended in the validation report (Hecker et al., 2011) and ultimately led to its inclusion in the refined Test Guideline (OECD TG 456, 2022). Our analyses, based on a more extensive and broadly sourced database support the refinement and indicate that the criteria for stringent assay conduct and the inclusion of statistical trend analysis proposed herein could help with interpretation in a regulatory context as they can take into account the data-driven characteristics of this assay and help identify truly robust effects that are appropriate for decision making rather than simply characterizing effects for "screening and prioritization purposes".

The H295R cells maintain the capacity for inducible production of steroid hormones. However, this is a rather sensitive system and thus stringent assay requirements are in place to ensure performance is meeting the necessary quality. These criteria include detailed specifications in the test guideline for cell culture including a custom media formulation as well as the requirement of freezing the cells after growth for five passages and use within no more than 10 passages. A QC plate is required to be run in tandem with test plates to provide a complete assessment of assay conditions. The test guideline provides plate layout and performance criteria for the QC plate. In addition to confirming passage numbers, our review began with the evaluation of QC plates to assess dynamic range requirements for conducting this assay. This evaluation revealed that several testing sites have wide distributions for induction and inhibition levels for E2 or T, some of which have means that do not meet the defined performance criteria. Scrutinizing QC plates is of relevance as limited dynamic ranges can compromise results obtained with test items. To robustly determine which results will have the greatest likelihood to be of biological relevance, a large dynamic range is critical.

For test item assessments, our results suggest that integration of trend testing and/or efficacy cutoffs would reduce the high rate of equivocal determinations and refine positive outcomes to only those that are robust and most likely to be biologically relevant. The trend testing allows for confirmation that effects are showing a statistically significant trend across concentrations. By applying a Jonckheere-Terpstra test no dose-response modeling is required and, as a non-parametric test, there are no assumption tests required either since the statistical approach is evaluating an ordered difference in medians. This simple approach agrees with the recent refinement of OECD TG 456 and can help provide more resolution to the criteria requiring two significant consecutive testing concentrations.

The database described herein provided insight into the impact that

4a



4b

Run analysis details for estradiol

Run	ANOVA Dunnett's p≤0.05 (µM)	Trend	Maximum Fold Change
1	10, 31.6, 100	$8x10^{-5}$	1.45
2	31.6, 100	$2.84x10^{-4}$	1.56

Overall outcome interpretation for estradiol

Run	Approach 1	Approach 2	Approach 3	Approach 4
1	Positive	Positive	Negative	Negative
2	Positive	Positive	Equivocal	Equivocal
Overall	**Positive**	**Positive**	**NA**	**NA**

4c

Run 1		Run 2		Run 3		Final Call
Scenario	*Decision*	*Scenario*	*Decision*	*Scenario*	*Decision*	
Negative	Confirm[a]	Negative	Stop			Negative
Negative	Confirm[a]	Positive/ Equivocal[c]	Refine[b]	Negative	Stop	Negative
Negative	Confirm[a]	Positive	Refine[b]	Positive	Stop	Positive
Negative	Confirm[a]	Positive	Refine[b]	Equivocal[c]	Stop	Weak Positive
Negative	Confirm[a]	Equivocal[c]	Refine[b]	Positive/ Equivocal[c]	Stop	Weak Positive
Equivocal[c]	Refine[b]	Negative	Confirm[a]	Negative	Stop	Negative
Equivocal[c]	Refine[b]	Negative	Confirm[a]	Positive/ Equivocal[c]	Stop	Weak Positive
Equivocalc	Refine[b]	Equivocal[c]	Confirm[a]	Equivocal[c]/ Positive	Stop	Weak Positive
Equivocalc	Refine[b]	Positive	Stop			Weak Positive
Positive	Refine[b]	Positive	Stop			Positive
Positive	Refine[b]	Negative	Confirm[a]	Positive	Stop	Positive

Fig. 4. Test item ID20 was tested in the H295R steroidogenesis assay (according to OECD TG 456, 2011) as part of its *in vitro* ED assessment.
Fig. 4a: The plot shows triplicate observations as boxplots for both runs for estradiol and testosterone. The estradiol increase resulted in a positive interpretation according to the original version of the test guideline (2011) as statistically significant increases in this hormone were observed in both runs at 31.6 and 100 µM
Fig. 4b: Tables summarize the statistical analyses and the interpretation derived from each run and analysis approach for estradiol; NA: Not Applicable was applied as the current decision matrix (OECD TG 456, 2022) does not account for the combination of results (negative, equivocal) and an additional run would have been required to conclude on the overall outcome.
Fig. 4c: Decision matrix taken from the current version of OECD TG 456 (2022). Arrows indicate the two possible final calls for test item ID20 if a third run had been performed.

low efficacy, but nonetheless statistically significant, outcomes in the steroidogenesis assay (based on the criteria of the original OECD TG 456 (2011)) could have on additional animal testing. Specifically, new Level 5 reproductive toxicity studies were triggered for 13/36 test items because of a positive or inconclusive outcome in the assay according to the original test guideline criteria. As can be seen in Fig. 3 and Tables 3 and 6 of those 13 compounds would have been classified as negative in the assay according to the recently refined OECD TG 456 (2022). Nevertheless, ~23,400 animals were used to address the relevance of the H295R assay outcomes for the 13 compounds in question.

To date only one compound included in our database has been officially concluded as non-ED for HH and two others are considered as non-ED in the DARs written by the RMS. In addition, a conclusion of non-ED for the EAS modalities has been proposed by the applicants for 8 of the other 10 test substances (no information is available at this time for the remaining two compounds for which the H295R data were voluntarily provided blindly for this project). In all cases these conclusions were based on either an absence of adversity in the extensive *in vivo* database available for each compound (including the new reproductive toxicity study) or an absence of a plausible link between the *in vitro* and *in vivo* observations.

The comprehensiveness of the *in vitro* and particularly the *in vivo* dataset used to determine whether an agrochemical is an ED was clearly demonstrated with the case study described herein. Test item ID20 was negative in all Level 2 assays except the H295R assay in which it was marginally positive for E2 induction. Specifically, the H295R assay had statistically significant increases in estradiol at two consecutive testing concentrations in two independent runs, with only the highest concentration in one run exceeding 1.5-fold. All *in vivo* bioassays including short- and long-term and reproductive studies previously conducted on the test item, as well as the extended one-generation reproductive toxicity assay (OECD TG 443, 2018b) that was triggered as a result of the positive H295R assay, revealed no adverse effects for the EAS-modalities. Test item ID20 has recently been concluded by EFSA as a non-ED for human health (internal information, EFSA conclusion not yet published) and is an example of where a weight of evidence approach with more stringent outcome interpretation criteria for the H295R assay could have been more effective in properly determining that the test item was in fact not of concern for ED at an earlier stage, negating the requirement for further evaluation in an *in vivo* reproductive toxicology test and thereby significantly reducing animal usage.

The refinement of OECD TG 456 (2022) not only includes the 1.5-fold threshold but also a new final call category of weak positive in the decision matrix. Currently, however, there is no clear guidance concerning what follow-up tests (if any) would be triggered in the case of a molecule being concluded as a weak positive. In the case of test item

Table 4
Summary of Findings Relevant to EAS-Mediated Effects from *in vivo* OECD Guideline Studies.

Species	Exposure	Highest Dose Administered	Effects Evaluated[a]	EAS Related Observations
rat	90 days/13 weeks (dietary studies; OECD TG 408; 2018c)	Study 1: 5000 ppm Study 2: 20,000 ppm Study 3: 20,000 ppm Study 4: 1500 ppm Study 5: 1200 ppm Study 6: 600 ppm	Endocrine organ weights and histopathology	No effect in the 6 studies. Study 3: ↑degree of diffuse vacuolation of zona fasciculata observed in males but not dose related and considered not treatment related.
rat	52 weeks (dietary studies)	Study 1: 1000 ppm Study 2: 1000 ppm	Endocrine organ weights and histopathology	No effect in either study
rat	104 weeks (dietary studies)	Study 1: 1000 ppm Study 2: 500 ppm Study 3: 1000 ppm Study 4: 2500 ppm	Endocrine organ weights and histopathology	Study 1: No effect; Minimal ↑ uterine sclerosis, not treatment-related (regarded as age-related finding) Study 2: No effect; Increased incidence of uterine endometrial stromal sarcoma (exact Cochran-Armitage test did not prove positive trend given high value in controls). Study 3: No Effect; Age related ↑uterine sclerosis Study 4: No effect; Slight ↑ pituitary adenoma in males but within HCD
rat	2-generation reproductive toxicity study (dietary study)	Study 1: 1000 ppm Study 2: 225 mg/kg/day (dietary adjusted)	Endocrine organ weights and histopathology. Reproductive efficiency parameters[b]	No effect in either study
rat	3-generation reproductive toxicity study (dietary study)	500 ppm	Endocrine organ weights and histopathology. Reproductive efficiency parameters[b]	No effect
rat	Extended one generation reproductive toxicity study (oral gavage; OECD TG 443)	200 mg/kg/day	Endocrine organ weights and histopathology. Reproductive efficiency parameters[b] (including sperm parameters). Pre and post weaning developmental landmarks	No ED related effects Microscopic findings related to changes in haematology parameters indicating methaemaglobinaemia were observed in males and females at all doses indicating general toxicity. Two F1 cohort 1a males in the top dose group had reduced cauda epididymal and tests weights with seminiferous tubule degeneration and reduced luminal sperm. These were considered incidental, commonly observed findings in young animals, which were not observed in F1 cohort 1b males.
mouse	8 weeks (dietary)	15,000 ppm	Endocrine organ weights and histopathology	No effect
mouse	78 weeks (dietary)	7000 ppm	Endocrine organ weights and histopathology	No effect: Increased amyloidosis in ovaries and adrenals not treatment-related (regarded as age-related finding)
mouse	104 weeks (dietary)	1000 ppm	Endocrine organ weights and histopathology	No effect
dog	60 days (dietary)	30,000 ppm	Endocrine organ weights and histopathology	No effect; ↓ relative adrenal weight and minimal – slight diffuse hypertrophy of zona glomerulosa considered to be due to severe toxicity at doses ≥3000 ppm
dog	18 weeks (oral, capsule)	Up to 1000 mg/kg/day	Endocrine organ weights and histopathology	No effect
dog	104 weeks (dietary)	1000 ppm	Endocrine organ weights and histopathology	No effect

[a] Gonadal histopathology comprises uterus, ovary, testis, epididymis, and histopathology; some studies also include prostate, vaginal, and mammary histopathology. Gonadal organ weights include uterus, ovary, testis, epididymis weights. Adrenal weights and histopathology were also included in the analyses.
[b] Reproductive efficiency parameters included mating, fertility, implantation sites, gestation indices, mean/median gestation length, litter parameters, sex ratio, and in some cases developmental effects or malformations.

ID20, the conduct of a third H295R assay would have been necessary to be able to conclude on its potential to interfere with steroidogenesis. Furthermore, depending on the results from this third run, the test item could be classified either as negative or as a weak positive according to the guideline's revised decision matrix. Although not a common occurrence in the extensive database described herein, a number of compounds were classed as weak positives according to the refined decision matrix (3 each for E and T and as an overall conclusion for 4 compounds). Furthermore, two compounds would warrant additional *in vitro* testing to finally conclude on steroidogenesis activity with potentially weak positive final calls. Clearly, further retrospective analyses will be required to fully understand the *in vitro* to *in vivo* extrapolation, particularly in cases of a weak positive outcome in the H295R assay.

In summary, the compendium of data compiled herein is the first of its kind to collate and assess H295R steroidogenesis data generated for regulatory purposes. This dataset was used to evaluate thousands of data points to allow data-driven determination of QC plate trends and assessment of different analysis approaches to determine potentially effective complementary criteria that can aid identification of robust positives and minimize equivocal outcomes. The H295R steroidogenesis assay is an important part of the Level 2 *in vitro* testing available to evaluate endocrine disrupting potential of chemicals and the analyses herein can help inform on how to best leverage data from this assay to provide information for regulatory purposes.

Credit author statement

H Tinwell: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Project administration. A Karmaus: Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. V Gaskell: Investigation, Writing – review & editing, C Gomes: Investigation, Writing – review & editing, C Grant: Writing – review & editing, T Holmes: Investigation, Writing – review & editing, A Jonas: Investigation, Writing – review & editing, S Kellum: Investigation, K Krüger: Investigation, Writing – review & editing, L Malley: Investigation, Writing – review & editing, S Melching-Kollmuss: Conceptualization, Writing – review & editing, O Mercier: Investigation, Writing – review & editing, H Pandya: Investigation,

Writing – review & editing, T Placke Investigation, Writing – review & editing, R Settivari: Investigation, Writing – review & editing, B De Waen: Investigation, Writing – review & editing.

Funding body information

This research was organized and financed by CropLife Europe.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

CropLife Europe organized and funded this work. We thank member companies that provided H295R testing data and ED conclusions for this analysis to be possible.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2023.105461.

References

Andersson, N., Arena, M., Auteri, D., Barmaz, S., Grignard, E., Kienzler, A., Lepper, P., Lostia, A.M., Munn, S., Parra Morte, J.M., Pellizzato, F., Tarazona, J., Terron, A., Van der Linden, S., 2018. Guidance for the identification of endocrine disruptors in thecontext of Regulations (EU) No 528/2012 and (EC) No 1107/2009. EFSA Journal 2018 16 (6), 135. https://doi.org/10.2903/j.efsa.2018.5311. ECHA-18-G-01-EN, 5311.

EPA, 2009a. Oppts 890.1550: steroidogenesis (human cell line H295R). Office of prevention, and toxic substances. Endocrine Disruptor Screening Program Test Guidelines. Docket #EPA-HQ-OPPT-2009-05676-0011.pdf.

EPA, 2009b. Oppts 890.1200: aromatase assay (human recombinant). Office of prevention, and toxic substances. Endocrine Disruptor Screening Program Test Guidelines. Docket # EPA-HQ-OPPT-2009-0576-0004.pdf.

EPA, 2011. Endocrine disruptor screening program for the 21st century: EDSP21 work plan; the incorporation of *in silico* models and *in vitro* high throughput assays in the endocrine disruptor screening program (EDSP) for prioritization and screening. https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf.

EPA, 2011b. OPPTS 890.1450: Pubertal Development and Thyroid Function in Intact Juvenile/Peripubertal Female Rats. Office of Prevention, and Toxic Substances. Docket #EPA-HQ-OPPT-2009-05676-0009.Pdf, Endocrine Disruptor Screening Program Test Guidelines.

Gazdar, A.F., Oie, H.K., Shackleton, C.H., Chen, T.R., Triche, T.J., Myers, C.E., Chrousos, G.P., Brennan, M.F., Stein, C.A., La Rocca, R.V., 1990. Establishment and characterization of a human adrenocortical carcinoma cell line that expresses multiple pathways of steroid biosynthesis. Cancer Res. 50 (17), 5488–5496.

Gracia, T., Hilscherova, K., Jones, P.D., Newsted, J.L., Zhang, X., Hecker, M., Higley, E. B., Sanderson, J.T., Yu, R.M., Wu, R.S., et al., 2006. The H295R system for evaluation of endocrine disrupting effects. Ecotoxicol. Environ. Saf. 65, 293–305.

Haggard, D.E., Karmaus, A.L., Martin, M.T., Judson, R.S., Setzer, R.W., Friedman, K.P., 2018. High-throughput H295R steroidogenesis assay: utility as an alternative and a statistical approach to characterize effects on steroidogenesis. Toxicol. Sci. 162, 509–534.

Haggard, D.E., Setzer, R.W., Judson, R.S., Friedman, K.P., 2019. Development of a prioritization method for chemical-mediated effects on steroidogenesis using an integrated statistical analysis of high throughput H295R data. Regul. Toxicol. Pharmacol. 109, 104510.

Hecker, M., Newsted, J.L., Murphy, M.B., Higley, E.B., Jones, P.D., Wu, R., Giesy, J.P., 2006. Human adrenocarcinoma (H295R) cells for rapid *in vitro* determination of effects on steroidogenesis: hormone production. Toxicol. Appl. Pharmacol. 217, 114–124.

Hecker, M., Hollert, H., Cooper, R., Vinggaard, A.M., Akahori, Y., Murphy, M., Nellemann, C., Higley, E., Newsted, J., Laskey, J., et al., 2011. The OECD validation program of the H295R steroidogenesis assay: phase 3. Final inter-laboratory validation study. Environ. Sci. Pollut. Res. Int. 18, 503–515.

Hilscherova, K., Jones, P.D., Gracia, T., Newsted, J.L., Zhang, X., Sanderson, J.T., Yu, R. M., Wu, R.S., Giesy, J.P., 2004. Assessment of the effects of chemicals on the expression of ten steroidogenic genes in the H295R cell line using real-time PCR. Toxicol. Sci. 81, 78–89.

Jacobs, M., 2021. Proposed TG 456 corrections. Presentation during OECD EDTA-AG 27. October 2021.

Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M.H., Huang, R.L., Rotroff, D.M., Filer, D.L., et al., 2015. Integrated model of chemical perturbations of a biological pathway using 18 *in vitro* high throughput screening assays for the estrogen receptor. Toxicol. Sci. 148, 137–154.

Karmaus, A.L., Toole, C.M., Filer, D.L., Lewis, K.C., Martin, M.T., 2016. High-throughput screening of chemical effects on steroidogenesis using H295r human adrenocortical carcinoma cells. Toxicol. Sci. 150, 323–332.

Kleinstreuer, N.C., Ceger, P., Watt, E.D., Martin, M., Houck, K., Browne, P., Thomas, R.S., Casey, W.M., Dix, D.J., Allen, D., et al., 2017. Development and validation of a computational model for androgen receptor activity. Chem. Res. Toxicol. 30, 946–964.

Maglich, J.M., Kuhn, M., Chapin, R.E., Pletcher, M.T., 2014. More than just hormones: H295R cells as predictors of reproductive toxicity. Reprod. Toxicol. 45, 77–86.

OECD, 2011. Test No. 416: Two-Generation Reproductive Toxicity Study, OECD Guidelines for the Testing of Chemicals. Section 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264070868-en.

OECD, 2011a. Test No. 456: H295R steroidogenesis assay, 2022. In: OECD Guidelines for the Testing of Chemicals, Section, 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264122642-en.

OECD, 2018b. Test No. 443: extended one-generation reproductive toxicity study. In: OECD Guidelines for the Testing of Chemicals, Section, 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264185371-en.

OECD, 2018a. Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption, OECD Series on Testing and Assessment. OECD Publishing, Paris. https://doi.org/10.1787/9789264304741-en.

OECD, 2018c. Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents, OECD Guidelines for the Testing of Chemicals. Section 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264070707-en.

OECD, 2020. Test No. 458: Stably Transfected Human Androgen Receptor Transcriptional Activation Assay for Detection of Androgenic Agonist and Antagonist Activity of Chemicals, OECD Guidelines for the Testing of Chemicals. Section 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264264366-en.

OECD, 2021. Test No. 455: Performance-Based Test Guideline for Stably Transfected Transactivation *in Vitro* Assays to Detect Estrogen Receptor Agonists and Antagonists, OECD Guidelines for the Testing of Chemicals, Section, 4. OECD Publishing, Paris. https://doi.org/10.1787/9789264265295-en.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.