Evaluating results from 55 extended one-generation reproductive toxicity studies under REACH

Final report of the EOGRTS review project

March 2023

Disclaimer

The information and views set out in this document do not necessarily reflect the position or opinion of the European Chemicals Agency or the Member States. The Agency does not guarantee the accuracy of the information included in the document. Neither the Agency nor any Member State nor any person acting on either of their behalf may be held liable for the use which may be made of the information contained therein. Statements made or information contained in the document are without prejudice to any further regulatory work that the Agency or Member States may initiate at a later stage.

Evaluating results from 55 extended one-generation reproductive toxicity studies under REACH: Final report of the EOGRTS review project

Reference: ECHA-23-R-04-EN
ISBN: 978-92-9468-262-8
Cat. Number: ED-05-23-079-EN-N
DOI: 10.2823/92503
Publ.date: March 2023
Language: EN

If you have questions or comments in relation to this document please send them (quote the reference and issue date) using the information request form. The information request form can be accessed via the Contact ECHA page at:
http://echa.europa.eu/contact
European Chemicals Agency
P.O. Box 400, FI-00121 Helsinki, Finland

Table of Contents

Authors

Project management: Ingo Bichlmaier and Virve Sihvola (ECHA)

Report coordination: Ulla Simanainen (ECHA)

Supervision: Hugues Kenigswald (ECHA)

ECHA Project Team: Niklas Andersson, Kati Hellsten, Hannele Huuskonen, Agnese Irkle, Outi Leppäranta, Kirsi Myöhänen, Laura Rossi

Evaluating experts nominated by competent authorities of the European Union Member States and the European Economic Area:

- Belgium: Sandrine Jouan (The Federal Public Service, FPS)
- Denmark: Marta Axelstad, Sofie Christiansen, Marie Louise Holmer, Ulla Hass (Technical University of Denmark, DTU)
- Finland: Marko Kuittinen, Eeva Rissanen, Tiina Suutari, Emma Tarvainen (Finnish Safety and Chemicals Agency, TUKES)
- France: Karine Angeli, Juliette Deweirdt (Agence nationale de sécurité sanitaire, ANSES)
- Germany: Olena Kucheryavenko, Esther Rosenthal, Gabriele Schöning, Achim Trubiroha (German Federal Institute for Risk Assessment, BfR), Wiebke Prutner (Federal Institute for Occupational Safety and Health, BAuA)
- The Netherlands: Wieneke Bil, Betty Hakkert, Joop de Knecht, Andre Muller, Petra van Kesteren, Jelle Vriend, Marjolijn Woutersen, Dion Zijtveld, Rob Vandebriel (National Institute for Public Health and the Environment, RIVM)
- Norway: Nina Landvik (Norwegian Environment Agency), Oddvar Myhre, Dag Markus Eide, Berit Granum, Birgitte Lindeman, Marcin Wojewodzic, Tim Hofer, Nur Duale, Camilla Svendsen, Ann-Karin Olsen, Espen Mariussen, Monica Andreassen, Hubert Dirven (Norwegian Institute of Public Health, NIPH)
- Spain: Pía Basaure García, Patricia García Hernández (Ministry of Health)
- Sweden: Charlotte Bergkvist (Swedish Chemicals Agency, KEMI)

Other contributors: Andrea Terron (European Food Safety Authority, EFSA), Martin Paparella and Kevin Crofton (external experts nominated by EFSA)[1]

[1] EFSA's team of experts only provided input on the DNT cohorts and not on any other aspects of this project.

List of annexes

List of figures

List of tables

Abbreviations

AGD	Anogenital distance
AGI	Anogenital index
CSA	Chemical safety assessment
CL	Corpora lutea
DNT	Developmental neurotoxicity
DIT	Developmental immunotoxicity
EOGRTS	Extended one-generation reproductive toxicity study/-ies
EU	European Union
F1	First filial generation
F1A	F1 pups of Cohort 1A
F1B	F1 pups of Cohort 1B
F2	Second filial generation
GD	Guidance document or gestation day
GLP	Good laboratory practice
IR	Information requirement
IUCLID	International uniform chemical information database
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LD	Lactation day or low-dose
LO(A)EL	Lowest-observed-(adverse-)effect level
mg/kg bw/day	Milligram per kilogram bodyweight per day
MS	Member State
MSCA	Member State competent authority
No.	Number
NO(A)EL	No-observed-(adverse-)effect level
P0	Parental animals used to be mated to produce F1
P1	Parental animals (adult F1) used to be mated to produce F2
PND	Postnatal day
REACH Regulation	Regulation (EC) No 1907/2006 as amended
SRBC	Sheep red blood cell
STOT RE	Specific target organ toxicity, repeated exposure
SVHC	Substance of very high concern
T3	Triiodothyronine
T4	Thyroxine
TDAR	T-cell dependent antibody response
TG	Test guideline
TSH	Thyroid stimulating hormone

Abstract

The focus of this project was to evaluate the performance of the extended one-generation reproductive toxicity study (EOGRTS) in terms of design, conduct, analysis, and reporting, and how the results support hazard assessment in the EU regulatory context.

To achieve this, the study summaries of REACH registration dossiers and the corresponding full study reports of 55 EOGRTS were evaluated.

The main findings highlighted issues related to identifying effect levels and reproductive toxicants/endocrine disruptors, using the F2 generation, selecting adequate dose levels, methodological aspects, reporting issues, and proficiency for conducting investigations.

While the EOGRTS was found to be generally effective in identifying substances of concern and the F2 generation was deemed useful, some methodological issues were identified. The selection of inadequate dose levels was regularly observed, and the developmental immunotoxicity (DIT) and developmental neurotoxicity (DNT) cohorts were particularly challenging in terms of test laboratory proficiency.

This report also includes two sections on recommendations to support test laboratories and sponsors to improve the design, conduct, analysis, and reporting of future EOGRTS.

Executive summary

This report presents the results of an assessment of 55 EOGRTS conducted with industrial chemical substances. Substances affecting fertility and development are strictly regulated in the EU, therefore it is essential to establish the performance of the EOGRTS aiming at investigating reproductive toxicological effects.

The objective of the project was to evaluate the performance of the EOGRTS in terms of design, conduct, analysis, and reporting and how the results support hazard assessment in the EU regulatory context.

To conduct the assessment, study summaries and full study reports of 55 EOGRTS from REACH registration dossiers were collected and made available to evaluators who used a standardised questionnaire to evaluate each study. For studies with developmental neurotoxicity cohorts, an additional team of experts nominated by the European Food Safety Authority (EFSA) evaluated the studies to provide further input. The results of each evaluation were presented and discussed at the substance-specific peer-review (calibration) meetings, where the outcome of the evaluation was agreed by all participating experts.

Overall, the EOGRTS was found to be effective in identifying chemicals with reproductive toxicity when it is conducted according to the OECD test guideline (TG) and EU regulation. However, 20 % of the studies conducted did not use adequate dose levels, hindering their ability to effectively identify potential hazards. This problem is not unique to EOGRTS. Despite some studies being performed with too low dose levels or deficiencies in conduct or reporting, around 30 % of the considered EOGRTS showed clear adverse effects on sexual function and fertility and/or development. It is likely that this percentage would be higher if all studies had appropriate dose levels, were well-conducted, and appropriately reported.

The most frequently observed methodological issues are presented with a specific focus on the developmental immuno- and neurotoxicity cohorts that appear to be particularly demanding in terms of proficiency.

This report also includes two sections on recommendations for methodologies and reporting to support test laboratories and sponsors in improving the design, conduct, analysis, and reporting of future EOGRT studies.

The challenges identified in evaluating complex studies such as EOGRTS suggest that the full study report should be attached in IUCLID.

1. Introduction

The EOGRTS was approved as OECD TG 443 in 2011[2] and became a REACH[3] information requirement in 2015, replacing the earlier information requirement for a two-generation reproductive study (OECD TG 416). ECHA's Guidance[4] was also updated with the test method in 2015.

OECD TG 443 describes the EOGRTS investigating reproductive endpoints that require the interaction of males with females, females with conceptus, and females with offspring and the F1 generation beyond sexual maturity. The TG provides a flexible study design under REACH, as the extension of Cohort 1B to produce F2 generation, Cohorts 2A and 2B, and Cohort 3 are only required if triggered based on available information. The criteria for triggers are described in REACH Annexes IX and X[3]. These criteria are explained in more details in the ECHA Guidance[4].

With respect to carcinogenicity, mutagenicity, reproductive, and specific organ toxicity, reproductive toxicity is the area for which the data generated under REACH and the related regulatory action by authorities have had the greatest impact on the classification of substances as shown in Karamertzanis et al. (2019)[5]. It is, therefore, crucial to examine the effectiveness of the EOGRTS in the European regulatory context.

The main objectives of the project are to:

- Evaluate the performance of the OECD TG 443 study as implemented in REACH, and its adequacy for providing information relevant for hazard assessment; and
- Evaluate specific aspects of the performance of available EOGRTS in REACH registration dossiers with respect to study design, conduct, and toxicological findings.

More specifically, the following aspects are addressed through the project:

- Toxicological aspects, in particular, considerations of adversity of findings and how these may contribute to effect-level setting, hazard classification, and SVHC/ED identification;
- Investigate how well the studies follow the specifications set out in ECHA's decisions as well as in OECD TG 443 and OECD guidance document (GD) 151[6], including dose-level selection; and
- Methodological aspects i.e. how the investigations are conducted, and how the results are analysed and reported.

This report describes the observed methodological issues with respect to designing and conducting the EOGRTS, as well as analysing its results and reporting of methodologies and

[2] OECD Guideline for the testing of chemicals 443, extended one-generation reproductive toxicity study: https://doi.org/10.1787/20745788
[3] Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4, as amended: https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:02006R1907-20220501
[4] ECHA's Guidance on information requirements and chemical safety assessment R.7a; R.7.6 Reproductive toxicity
[5] Karamertzanis et al. *The impact on classifications for carcinogenicity, mutagenicity, reproductive and specific target organ toxicity after repeated exposure in the first ten years of the REACH regulation.* Regulatory Toxicology and Pharmacology, Volume 106 (2019), 303-315: https://doi.org/10.1016/j.yrtph.2019.05.003
[6] OECD Guidance Document 151 in support of OECD Test Guideline 443 on an Extended One Generation Reproductive Toxicity Study: https://www.oecd.org/chemicalsafety/testing/seriesontestingandassessmenttestingforhumanhealth.htm

results. It is important to consider that the impact of these issues on the acceptability and usability of the results can vary. For instance, some investigations may have minor deficiencies that cannot be rectified, but only affect the validity of certain parameters, without impeding the study's applicability for regulatory purposes. Similarly, reporting and statistical analysis shortcomings can typically be remedied. However, significant deficiencies such as the selection of too low dose levels or the omission of mandatory investigations may render a study invalid and a re-run of the EOGRTS might be needed to meet regulatory requirements. This report supports registrants and test laboratories to provide adequate studies to ensure that the assessment of reproductive hazards can be effectively performed.

In addition to the main project covered in this report, there are three complementary analyses (referred to as "satellite-projects"), each with specific research questions and objectives. These projects focus on thyroid hormone measurements (with participants from France, Germany and ECHA), nipple retention and anogenital distance (Denmark, France, Sweden and ECHA), and follicular/corpora lutea count (France and ECHA). The satellite-projects are expected to conclude by the end of 2023.

2. Methodology

This project analysed EOGRTS requested by ECHA in the context of both dossier and substance evaluations. The analyses and conclusions of this project are based on the evaluation of EOGRTS results, from full EOGRTS reports and robust study summaries in the IUCLID registration dossiers provided to ECHA.

For this report, the results of 55 studies were evaluated. The substances, for which the EOGRTS results were evaluated, are listed in Table 1. These included results from 24 DNT cohorts, 14 DIT cohorts, and 23 studies with extensions of Cohort 1B to produce F2 generations.

The evaluation work was conducted by a team of scientific and regulatory experts with expertise in reproductive and developmental toxicology, endocrine toxicology, neurotoxicology, immunotoxicology, and statistics. These experts were from ECHA and the competent authorities of several EU Member States and European Economic Area countries, namely Belgium, Denmark, Finland, France, Germany, the Netherlands, Norway, Spain, and Sweden. Additionally, experts from the European Food Safety Authority (EFSA) provided evaluations of the DNT cohorts.

This project only evaluated a sample of EOGRTS including information on dose-range finding studies. The sample was selected based on the passed deadlines for providing requested data and availability of finalised full study reports. The evaluation did not involve a systematic weight-of-evidence assessment to compare the results of EOGRTS with those of other toxicological studies, such as repeated dose or developmental toxicity studies.

During calibration meetings, all participating experts reported their evaluation results, and the outcomes were agreed upon by consensus. The evaluation outcomes were reported in a standardised questionnaire (see the annex).

The calibration meetings aimed to ensure consistency, and all experts were invited to attend. Interested experts participated based on their availability. All calibration meetings were chaired and notes were taken by ECHA staff. To ease the workload, some cases were concluded in written procedure and not discussed at the calibration meetings if they met the following requirements: highest dose was limit dose and no adverse effects were observed at any dose level.

The annex contains an overview table of the investigations to be conducted in the EOGRTS according to OECD TG 443[2], OECD GD 151[6] and ECHA Guidance[4]. This table can serve as a reference guide.

A draft interim report was circulated on 9 May 2022, to volunteering stakeholder observers[7], test laboratories[8], experts from participating competent authorities, the European Commission including the Joint Research Centre, and EFSA for comments until 30 May 2022. After incorporating the feedback, an updated version of the report was shared with the commenting parties on 7 October 2022.

Originally, the project aimed to evaluate EOGRTS for more than 100 substances. However, after analysing around 40 studies, the evaluators realised that continuing with further cases would not add much value in identifying new issues. Consequently, the project was limited to evaluating 55 EOGRTS.

On 13 February 2023, the draft final report was distributed to volunteering stakeholder observers, test laboratories, experts from participating competent authorities, the European Commission including the Joint Research Centre, and EFSA for comments by 27 February 2023. After taking the comments into account, the report was revised and published as this final report.

Table 1: Substances for which the EOGRTS were evaluated, including the study design.

EC number	*F2*	*DNT*	*DIT*	*EC number*	*F2*	*DNT*	*DIT*
200-657-5	Yes	Yes	Yes	210-483-1	No	Yes	Yes
200-830-5	No	No	No	212-344-0	No	Yes	No
200-843-6	Yes	Yes	No	214-946-9	Yes	No	No
201-074-9	Yes	No	No	216-133-4	Yes	No	No
201-116-6	No	No	No	221-110-7	Yes	No	No
201-289-8	Yes	Yes	Yes	221-975-0	No	No	No
201-645-2	No	Yes	No	228-250-8	Yes	Yes	No
201-939-0	Yes	No	No	229-962-1	No	Yes	Yes
202-013-9	Yes	No	No	231-391-8	No	No	No
202-213-6	Yes	Yes	Yes	233-788-1	No	No	No
202-319-2	Yes	No	No	245-509-0	No	No	No
202-708-7	No	Yes	No	246-807-3	No	No	No
202-785-7	Yes	Yes	Yes	256-032-2	No	No	No
203-002-1	Yes	No	No	262-872-0	Yes	Yes	Yes
203-268-9	Yes	No	No	270-128-1	Yes	Yes	No
203-614-9	No	No	No	428-310-5	No	No	No
203-615-4	Yes	Yes	Yes	500-097-4	Yes	Yes	Yes
203-737-8	Yes	No	No	500-130-2	No	Yes	Yes
203-815-1	No	No	No	614-503-3	No	No	No
203-868-0	No	Yes	Yes	618-561-0	No	No	No
204-104-9	No	No	No	620-174-7	No	No	No
204-127-4	No	No	No	700-960-7	No	Yes	Yes
204-289-6	No	Yes	No	906-170-0	Yes	Yes	Yes
204-626-7	Yes	No	No	908-918-1	No	No	No
204-662-3	No	No	No	915-589-8	No	Yes	No
204-709-8	No	No	No	915-730-3	No	No	No
205-743-6	Yes	Yes	Yes	940-592-6	No	Yes	No
205-769-8	No	Yes	No				

Discussions with the test laboratories[8] responsible for carrying out the majority of the EOGRTS offered valuable practical knowledge pertaining to the study's planning, design, implementation, and reporting, as well as the analysis and assessment of its findings.

[7] Stakeholder observers are European Environmental Bureau (EEB), the Organisation for Economic Co-operation and Development (OECD), People for the Ethical Treatment of Animals (PETA), Cruelty Free Europe (CFE), Human Society International (HSI), The European Chemical Industry Council (Cefic), and Eurometaux.
[8] Charles River Laboratories and Labcorp (formerly Covance).

3. Main findings

3.1 The studies support the identification of substances of very high concern

Although some studies were performed with too low dose levels and/or deficient conduct and reporting, around 30 % of the EOGRTS showed clear adverse effects on sexual function, fertility and/or development. It is likely that this percentage would be higher if all studies had used appropriate dose levels and been conducted properly.

Classification and labelling. For 14 cases, the results, as concluded by the project evaluating experts, support classification as reproductive toxicant category 2 (Repr. 2)[9] and for 18 cases as reproductive toxicant category 1B (Repr. 1B)[10]. Typical effects that were considered for classification for sexual function and fertility included effects on fertility index, sperm, seminiferous tubules, oestrus cycle, follicle counts, preimplantation loss/number of implantations, sexual maturation, dystocia, testis, and epididymis. For development, these included effects on pup body weights, pre- and postnatal loss including stillbirths, hydrocephaly, developmental neurotoxicity (e.g. brain morphometry, motor activity) and developmental immunotoxicity.

The results of 14 EOGRTS were also considered to support a STOT RE classification[11] based on effects on liver, kidney, heart, brain, thyroid, white blood cells, spleen, and thymus, for example. 13 of these studies also supported classification for reproductive toxicity (5 for Repr. 2 and 8 for Repr. 1B). However, the observed effects in adults that may contribute to STOT RE classification were either observed at higher dose levels than those causing reproductive toxicity, or the toxicity in adult animals was not considered to question the classification for reproductive toxicity as the effects on reproduction were significant, and it was not possible to unequivocally demonstrate that the effects were secondary to maternal toxicity.

11 of the EOGRTS failed to demonstrate clear effects and were therefore considered inconclusive for classification and labelling because treatment groups were exposed to too low doses. The results of 3 of these might contribute to Repr. 2 classification.

SVHC/ED identification. According to Article 57(c) of REACH, substances meeting the criteria for classification in the hazard class reproductive toxicity category 1A or 1B, adverse effects on sexual function and fertility or on development, can be identified as substances of very high concern (SVHCs).

For 18 substances, the results support classification as Repr. 1B for sexual function and fertility and/or development (see above). Therefore, these substances could be considered for SVHC identification according to Article 57(c) if the substances are classified for reproductive toxicity.

In addition, findings relevant for identifying endocrine disruptors can contribute to SVHC identification according to Article 57(f). Criteria for substances with endocrine disrupting properties, require scientific evidence of probable serious effects to human health which give rise to an equivalent level of concern to those of other substances that meet the criteria for classification in the hazard class carcinogenicity, germ cell mutagenicity, and/or reproductive

[9] 10 x Repr. 2/f, 4 x Repr. 2/fd, and 3 x Repr. 2/d (results on the same substance may be considered to support classification for both development as well as sexual function and fertility)
[10] 10 x Repr. 1B/FD, 5 x Repr. 1B/D, and 3 x Repr. 1B/F (results on the same substance may be considered to support classification for both development as well as sexual function and fertility)
[11] 2 x STOT RE 1 and 12 x STOT RE 2

toxicity category 1A or 1B.

The EFSA-ECHA Guidance for the identification of endocrine disruptors[12] describes detailed parameters within EOGRTS that are sensitive to or diagnostic of oestrogenic, androgenic, thyroidal and steroidogenic modalities that provide information on endocrine activity and adversity which may be endocrine-related (refer to Table 14 in the EFSA-ECHA Guidance[12]). To identify endocrine disruptors, all available information for the substance including the EOGRTS, must be evaluated using a weight of evidence approach.

At least 16 EOGRTS provide information relevant for the identification of the respective substances as endocrine disruptors based on the thyroid-related findings (most often) but also on sexual maturation, anogenital distance and nipple retention.

3.2 The F2 generation can be effectively used to identify new effects and confirm findings

Cohort 1B was extended in 23 studies to produce the F2 generation. Triggering was based on column 2 criteria in Annex IX/X, Section 8.7.3 of the REACH Regulation for dossier evaluation. However, in substance evaluation, the inclusion criteria for the extension of Cohort 1B to generate F2 may be different. Additionally, in five of these studies, in-study triggers were used to justify the extension of Cohort 1B for producing the F2 generation, for example, to follow OECD GD 117[13] advice on internal triggers in the United States and Canada.

To fully benefit from the F2 data, identical investigations in the F1 and F2 generations should be conducted. In some cases, thyroid hormone measurements were missing in F2 pups, for example. Clarifying the triggering concerns with F2 also suffered from issues with dose level setting. Nevertheless, four studies showed effects only in the P1/F2 generations resulting in lower NOAELs in two of the studies based on effects observed in the F2 generation. Higher sensitivity in P1/F2 compared to P0/F1 also contributed to NOAEL setting in seven studies. In 15 studies, findings seen in P0/F1 were confirmed with P1/F2 animals, strengthening the weight of evidence in an evaluation and demonstrating clear value of the F2 generation for hazard assessment. Finally, Cohort 1B was extended by registrants to investigate equivocal findings observed in P0/F1 in two cases, but the results failed to confirm the identified concerns for reduced fertility and early deaths.

Effects observed in P1/F2, not in P0/F1. In four studies, the following effects were only observed in P1/F2 and not in P0/F1 animals:

- Reduced fertility index in P1;
- Reduced number of implantation sites in P1;
- Reduced anogenital distance in F2; and
- Increased percentage of dead pups in F2.

The effects observed in F2 contributed to lower NOAELs in two of these studies.

Higher sensitivity in P1/F2 compared to P0/F1. In seven studies, the following effects were

[12] Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009: https://echa.europa.eu/en/-/guidance-on-identifying-endocrine-disruptors-published
[13] OECD Guidance document 117 on the current implementation of internal triggers in Test Guideline 443 for an extended one generation reproductive toxicity study, in the United States and Canada: https://www.oecd.org/chemicalsafety/testing/48516094.pdf

observed at a lower dose and/or higher severity in P1/F2 compared to P0/F1:

- Hydrocephaly at lower dose in F2 than F1;
- Effect on gestation length at a lower dose in P1 than P0, and effect on gestation length, post-implantation survival, and live birth index more severe in P1/F2;
- Nipple retention at lower dose and more severe in F2;
- Increased postnatal loss on PND 4 at lower dose with statistical significance in F2 compared to F1 where this was seen at higher dose without statistical significance;
- Post-implantation loss at lower dose in P1 than in P0, however, other developmental effects at same doses in F1 and F2;
- Decreased litter size at lower dose in P1/F2 compared to P0/F1; and
- Reduced fertility index at a lower dose in P1 compared to P0 animals.

The effects observed in P1/F2 contributed to NOAEL setting in seven studies.

Findings in P1/F2 confirming those in P0/F1. In 15 studies, the following effects observed in P0/F1 were confirmed in P1/F2:

- Hydrocephaly;
- Increased incidence of dystocia;
- Nipple retention;
- Reduced number of implantation sites;
- Reduced pup body weights;
- Postnatal loss;
- Lower brain weight;
- Reduced T4; and
- Effects on follicle counts in F1 associated with lower fertility index in F1B.

Findings in P1/F2 to investigate unclear findings in P0/F1. In two studies, where Cohort 1B was extended by registrants to investigate equivocal findings observed in P0/F1, the identified concerns for reduced fertility and early deaths in P0 were not confirmed in P1.

3.3 Types of effects driving the NOAELs and LOAELs

The most common effect leading to the lowest NOAEL for sexual function and fertility is reduced number of implantations in the treated groups compared to controls. For developmental toxicity, the most common finding leading to a lowest NOAEL was reduced post-implantation survival, including postnatal mortality.

For eight studies, NOAEL values could not be identified due to effects occurring already at the lowest dose level tested in the EOGRTS. No precise NOAEL value could be defined in 12 studies because there are no adverse effects on reproduction or general systemic toxicity at the highest

dose level. In 5 of these 12 studies, the study was conducted up to the limit dose of 1 000 mg/kg bw/day and therefore these were considered to be not reprotoxic.

In the remaining seven studies the top dose was below the limit dose and the studies were deemed inconclusive for classification and labelling by evaluators. It is also noted that in 5 of these 7 studies, the top doses are below 500 mg/kg bw/day, i.e. significantly below the limit dose. The top doses in these five studies are between 15 and 400 mg/kg bw/day.

Comparison of the LOAEL for systemic and reproductive toxicity. In 13 studies, the lowest LOAEL for reproductive toxicity, meaning toxicity on sexual function and fertility, and development, was higher than for systemic toxicity while in 15 studies, the lowest LOAEL for reproductive toxicity was lower than that for the systemic toxicity. In 15 studies, the lowest LOAELs were the same for reproductive toxicity and systemic toxicity and in 12 studies no effects were reported at the highest dose tested (Figure 1).

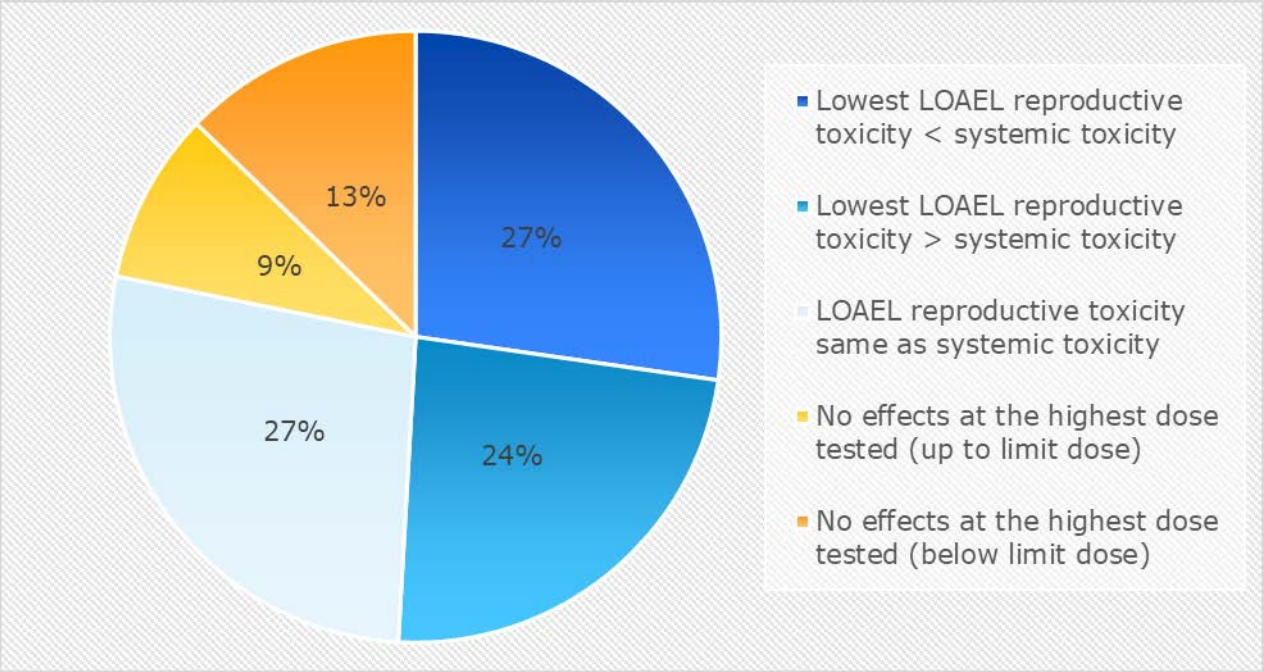

Figure 1: Percentage of cases (total of 55) where the lowest LOAEL for reproductive toxicity was higher, lower or the same as for the systemic toxicity or where no effects were reported at the highest dose tested.

Comparison of the LOAELs for developmental toxicity and sexual function/fertility. In eight studies, the lowest LOAEL for the effects on sexual function and fertility was lower than for the effects on development, while for 20 cases the LOAEL for effects on development was lower than for sexual function and fertility. In 11 studies, the lowest LOAELs were the same for sexual function and fertility and for development, and in 16 studies no effects on reproductive parameters were reported at the highest dose tested (Figure 2).

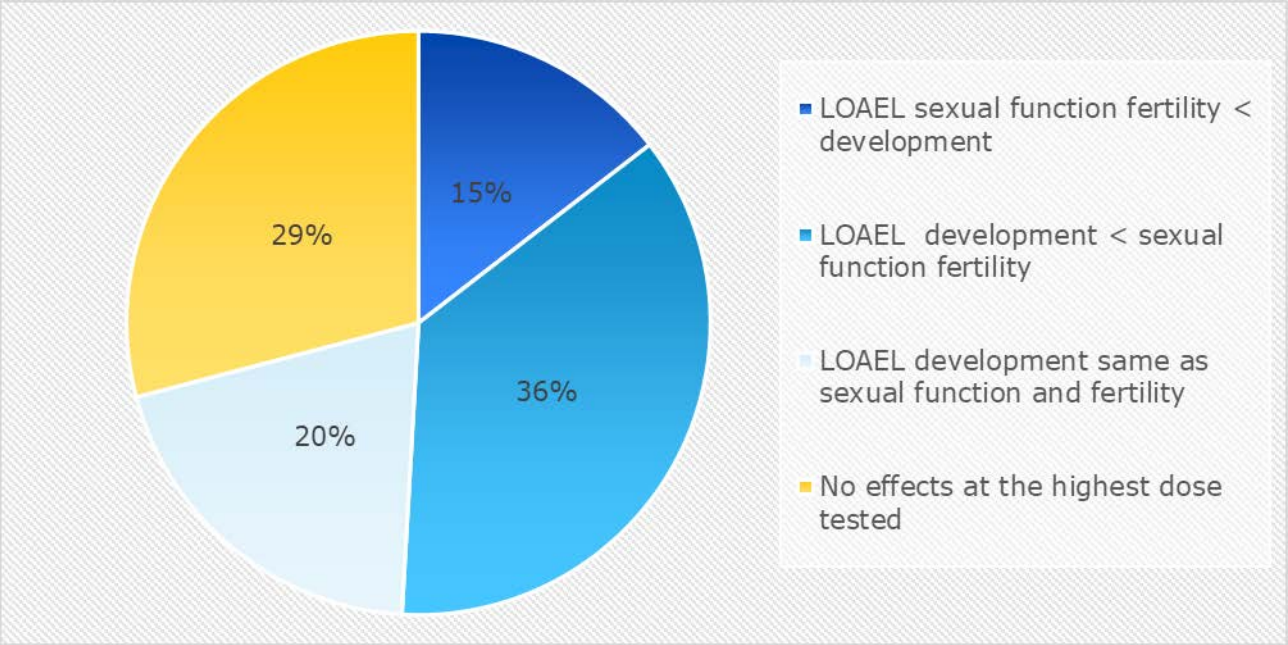

Figure 2: Percentage of cases (total of 55) where the lowest LOAEL for the effects on sexual function and fertility were higher, lower or the same as the lowest LOAEL for the effects on development or where no effects were reported at the highest dose on reproductive parameters.

3.4 Insufficient dose levels hampers hazard identification in 20 % of the studies

From the EOGRTS evaluated within the project, 11 out of 55 (20 %) of the studies were performed at doses that were too low to draw conclusions about the endpoint or adequately address concerns. The issue is not unique to EOGRTS, as underdosing is present in other toxicological studies submitted for REACH purposes[14]. Consequently, ECHA published advice[15] on selecting appropriate dose levels for reproductive toxicity studies in January 2022, to support assessing the safety of chemicals. The need to select adequate dose levels is also clarified in the amended REACH annexes[16].

The main objective of the toxicity studies is to gather information on the intrinsic toxicological properties of the substances. In the regulatory context, data generated from these studies should support the identification of hazards and assessing the risks associated with their use, ultimately ensuring adequate protection of human health. Therefore, as explicitly stated in the REACH Annexes on information requirements (as amended by Commission Regulation (EU) 2021/979 of 17 June 2021): "*Where a test method offers flexibility in the study design, for example in relation to the choice of dose levels, the chosen study design shall ensure that the data generated are adequate for hazard identification and risk assessment. To this end, testing shall be performed at appropriately high dose levels*.".

In the EU regulatory context, the data generated by an EOGRTS should allow a conclusion to be made on classification and labelling and on whether the substance meets the criteria for a substance of very high concern regarding endocrine disruption according to Article 57(f) of REACH. Under REACH, the EOGRTS is requested to provide information mainly on the sexual function and fertility, but it may also provide information on developmental toxicity and

[14] https://echa.europa.eu/-/new-advice-for-determining-dose-levels-in-toxicity-testing
[15] https://echa.europa.eu/documents/10162/17220/211221_echa_advice_dose__repro_en.pdf/27159fb1-c31c-78a2-bdef-8f423f2b6568?t=1640082455275
[16] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0979&from=EN

endocrine disrupting properties as well as effects on or through lactation and other toxicity. Priority is, however, given to fertility effects, meaning that the study should be designed to ensure adequate assessment of the effects on sexual function and fertility and the dose levels should not be reduced to get enough offspring to assess developmental toxicity.

The aim is that the highest dose level should be sufficiently high to conclude if a substance is a reproductive toxicant or not, and the lowest dose level should show no toxicity to identify a NOAEL. For 11 studies, it was not possible to conclude if the substances exert reproductive toxic properties warranting a classification as Repr. 1B due to an insufficient dose level selection. If the top-dose is well below the limit dose of 1 000 mg/kg bw/day and if only minimal or even no toxicity is observed, the studies have limited or no value for hazard identification. Furthermore, both sexes should be taken into consideration in dose-level selection to ensure that reproductive toxicity in either sex is not overlooked[17]. In practice, the less sensitive sex should be tested at higher doses than the more sensitive sex[18]. In addition to the 20 % of cases that failed adequate dose-level selection, 3 studies (5 %) had adequate dose-level setting for only one sex.

Dose-level selection should be based on existing information, not theoretical considerations. If the existing studies are not sufficient to inform on dose-level selection, a dose range finder including pregnant animals and pups is crucial to minimise the risk of unforeseen toxicity in the main study and also to demonstrate the aim of meeting requirements for dose-level setting.

3.5 Missing investigations, deviations in reporting, and demonstrating proficiency

The EOGRTS serves as a definitive study for sexual function and fertility by providing information on all phases of the reproductive cycle, as well as on development, growth, survival, and functional endpoints of offspring. However, failure to perform obligatory investigations listed in OECD TG 443 or deviation from triggered measurements requested in ECHA's final decision can impede the ability of a study to address its regulatory purpose.

The most frequently missing investigation seen during the project was the triggered evaluation of Cohort 1B histopathology. The histopathology investigation in Cohort 1B should be conducted to clarify equivocal results seen in Cohort 1A or in cases of suspected reproductive and/or endocrine toxicants[19], as clarified in ECHA's final decision. Triggering of the EOGRTS at Annex IX can be based on previous concern of reproductive or endocrine parameters and in those cases histopathology of Cohort 1B should be investigated. Similarly, as in case of equivocal results, unclarified concerns due to missing histopathological investigations do not allow comprehensive hazard assessment and definitive conclusions to be drawn.

For the histopathology investigations in general, if primary investigation is done only with control and high-dose animals, also low-dose and mid-dose groups should be evaluated in case treatment related effects are observed. Results obtained only from control animals and treatment related effects from high-dose animals will hinder NOAEL setting.

Furthermore, the absence of obligatory investigations such as thyroid hormone analysis in F2 pups on postnatal day 22 was frequently noted and two cases were lacking splenocyte subpopulation analysis in Cohort 1A. These parameters are crucial to conclude on the intrinsic properties of a substance and any deviation from these parameters should be justified and

[17] ECHA's Guidance on information requirements and chemical safety assessment R.7a, R.7.6.2.2.3: https://echa.europa.eu/documents/10162/17224/information_requirements_r7a_en.pdf/e4a2a18f-a2bd-4a04-ac6d-0ea425b2567f?t=1500286622893
[18] During cohabitation (mating), the dose levels need to be adjusted to those of the more sensitive sex to avoid severe suffering and death.
[19] OECD TG 443, paragraph 67 (see footnote 2)

reported, to allow proper interpretation and independent evaluation of the results. The objectives of the OECD TG 443 study design are elaborated in ECHA's Guidance on Information Requirements and Chemical Safety Assessment[4] and in the OECD Guidance Document 151[6].

Even if all parameters are not equally critical for the overall usefulness and acceptability of the study, missing investigations without justification always raise concerns on reliability.

Careful reporting of measured parameters is an essential step for regulatory acceptance. Evaluating experts should be able to perform an independent assessment based on the information available. The usual standard reporting in robust study summaries in IUCLID format may be challenging for complex studies such as the EOGRTS. Nevertheless, it is important to report all measured parameters and results in a tabular form, and if possible, attach the full study report to the technical IUCLID dossier. Additionally, presenting data in graphical form (in addition to tables) can help the interpretation of results, e.g. body weight or auditory startle.

Historical control data was not consistently provided even where there is a concern that the concurrent control group does not seem to represent typical values for certain parameters. It can aid in interpretation of findings such as anogenital distance, spermatogenesis, nipple retention, ovary follicle count, thyroid hormone concentrations, as well as immunotoxicity (splenocyte subpopulation, TDAR), and neurotoxicity, parameters that are not part of any other study design. Furthermore, it is good practice to include historical control data if the observed finding is interpreted as non adverse in reference to historical control data and observed change from controls (at least at the top dose) is greater than 10 % or when statistical significance is reached (at least at the top dose). Historical control data can only be considered valid if the methodology used for measuring the background is appropriate and reliable. It is recommended to follow paragraph 67 of OECD GD 43[20] to provide historical control data.

To succeed in performing complex studies such as OECD TG 443, the test laboratory should demonstrate proficiency. The full study report should contain some information on the validation proficiency of the methodology in the laboratory such as historical control data and positive control data to demonstrate that the laboratories have technical proficiency to generate reliable and reproducible data and that the methodology is sensitive enough to detect changes in the evaluated parameters. This was not done in any of the 55 evaluated EOGRTS reports. When commercially available kits or devices are used, these should be identified, as well as the possible computational software used. Furthermore, adequately trained staff should carry out the investigations and correct statistical analyses should be applied. The information should allow evaluators to conclude if the laboratory has sufficient experience and proficiency in conducting the assays.

3.6 Methodological deficiencies made the Developmental Immunotoxicity Cohort frequently inconclusive

Results from measuring functional immune parameters such as suppression or enhancement of immune response are essential when evaluating developmental immunotoxicity.

During the project, 14 DIT cohorts were evaluated and 4 TDAR assays (29 % of studies with DIT cohort) gave positive results for immunomodulation. In one positive case, the DIT cohort was in-study triggered, but no justification was reported. For the other positive DIT cases, one was triggered based on thymus atrophy and two were triggered based on an estrogenic mode of action observed in existing studies.

[20] OECD Guidance Document on mammalian reproductive toxicity testing and assessment: https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)16&doclanguage=en

Altogether, there were 6 cases triggered based on hormonal mode of action.

The mandatory measurement of splenic lymphocyte subpopulations in Cohort 1A animals did not consistently show effects in TDAR-positive studies. However, in the case triggered by thymus atrophy, adverse effects in the splenic lymphocyte subpopulation were also observed. One of the 14 cases suffered from too low dose level selection, and the DIT methodology was compromised due to a high number of non-responders in the control group.

For most studies, very wide coefficients of variation were observed for the TDAR (Cohort 3), making it difficult to detect dose responses. Therefore, such studies should be considered as inconclusive instead of negative. Other deficiencies in TDAR methodology included a lack of positive control and historical control data. Excessive variability, compared to other test laboratories or published data, suggests that a laboratory may not have adequate control of test conditions. Together, a lack of positive and historical control data increases the possibility of false negative findings.

Achieving a dose response with sufficiently high dose level selection is crucial. Furthermore, interpreting the outcome in the context of other conventional parameters, such as organ weights, histopathology and clinical chemistry, is essential. Multiple biomarkers need to be evaluated together in a weight of evidence assessment to identify hazards to the immune system.

3.7 Methodological deficiencies hindered the interpretation of most Developmental Neurotoxicity Cohorts

During the project, 24 DNT cohorts were evaluated. The inclusion of the DNT cohort increases the complexity of the EOGRTS, which most test laboratories were not prepared for. As a result, deficiencies were frequently found that hindered the interpretation of the results.

In all studies, there was a lack of positive and historical control data for mandatory investigations[21]. The lack of positive control data makes it difficult to determine whether the test laboratory is capable of detecting any aberrant effect following developmental exposure to a chemical. It also provides no information on the dynamic range of the test system, which would allow a better interpretation of the degree of change that is biologically relevant.

Additionally, the lack of historical control data makes it impossible to determine whether the control data in the study report are consistent with previous measurements by the test laboratory, especially because the statistical power in DNT investigations is rather low due to the low number of animals per group. Excessive variability in comparison to other test laboratories or published data suggests that a test laboratory may not have adequate control of test conditions. Together, the lack of positive and historical control data increases the possibility that negative results may be false negative findings.

Another reoccurring deficiency was inadequate statistical analysis. Most commonly, sex was not included in the statistical model. As a result, no conclusions can be made about sex-specific effects, or a lack thereof. Ignoring small effects with different directions in males and females while not performing correct statistics should be considered a relevant limitation. To address this, it is recommended to always perform both a combined and a sex-specific analysis of the results to avoid overlooking sex-specific events, whereas a combined analysis would result in a higher statistical power. Statistical analysis deficiencies may lead to uncertainty in NOAEL values or even prevent a proper conclusion on the effects being drawn. In cases where inadequate statistical analysis has been performed, a statistical re-analysis may be necessary to

[21] OECD TG 443, paragraph 80 (see footnote 2)

enable proper conclusions to be drawn.

In the cases evaluated during the project, brain weight and brain morphometry were observed to be the most sensitive parameters measured. In general, brain adversity is detected based on changes in brain weight, dimensions, morphometry and histopathology.

Overall, the independent evaluation of the data provided posed several challenges as the conclusions drawn from these studies could not be verified due to numerous methodological inadequacies that were either test-specific or common across different tests.

In addition, some studies suffered from limited reporting, even when considering the full study report. Detailed reporting is crucial, including information on the equipment used, control of extraneous experimental factors, the time of day testing is performed, as well as unbiased testing of animals.

Finally, the project findings call for significantly improved reporting, given that none of the current OECD TGs referred to in EU regulations and OECD/EU guidance documents relevant to developmental neurotoxicology provide exact requirements for reporting methods and results. It is important to report all observations, including normal and random observations, to show that observations were made with sufficient sensitivity.

4. Recommendations on study design

The project had a focus on evaluating the extent to which the studies provided by registrants followed the specifications set out in ECHA's decisions.

These decisions define the study design requirements and provide clarity on various aspects, such as the test material specification, route of administration, species (and the strain, if needed), requirements for dose level setting, premating exposure duration, expansion of the basic study design (specific cohorts to be included in the EOGRTS) and justification for triggering.

Building on the observations, recommendations are provided to help registrants and laboratories to comply with ECHA decisions and avoid potential issues.

4.1 Good laboratory practice

New studies conducted for REACH must comply with the principles of good laboratory practice (GLP). For the studies considered in this project, all laboratories had an adequate GLP status. In one case, however, some results raised doubts on the conduct of the study: the dose-range finding study included results for ovary weights in male animals and the main EOGRTS showed higher testosterone levels in females compared to males.

Registrants should carefully select the laboratory that carries out tests for them. The test facility must be regularly inspected within the OECD GLP monitoring programme by a national GLP monitoring authority. The test facility must also have a GLP certification for the area of expertise relevant for the particular test – in this case, for reproductive studies including developmental neurotoxicity and immunotoxicity.

4.2 Route of administration

Under REACH, testing for reproductive toxicity aims to maximise internal exposure of the substance. The oral route is preferred for solids and liquids, and the inhalation route is applicable for gases[3]. In all cases, registrants complied with the requested route of exposure. Oral administration was used in 52 studies (36 x gavage, 12 x dietary and 4 x drinking water), and inhalation exposure in three studies (2 x whole body and 1 x nose-only).

4.3 Species / strain

Rats are the preferred species, and criteria and recommendations given in OECD TG 443 refer to rats. The rat strain is specified in the request of ECHA's decision if there is evidence that a certain strain is more sensitive than others or if relevant findings related to reproductive toxicity were observed in studies with a certain strain. In three decisions, the rat strain was specified as Wistar and these strains were also used in these studies. All of ECHA's decisions requested testing on rats and all studies were conducted with either Wistar Han or Sprague-Dawley rats. Overall, Wistar Han rats were used in 33 and Sprague-Dawley rats in 22 studies.

The most sensitive rat strain with respect to reproductive effects should be used in the EOGRTS. For example, if existing data shows reproductive effects in Wistar Han rats, then this strain should be used in preliminary studies (i.e. dose range finding studies) and the EOGRTS.

4.4 Premating exposure duration

According to ECHA Guidance[4], the default premating exposure duration for P0 animals is 10 weeks. If the extension of Cohort 1B is requested, the premating exposure duration can be reduced to at least two weeks if the substance does not show delayed steady state kinetics.

In 43 studies, the premating exposure duration was 10 weeks and in 12 studies at least two weeks. All conducted studies complied with the requested premating exposure duration. However, Cohort 1B was extended in eight studies although not requested in the decision. The respective decisions requested a 10-week premating exposure duration, and the registrant did not evaluate if it could have been reduced to at least two weeks. Of course, such decisions can only be made if the decision to extend Cohort 1B is made before the study is initiated.

If, based on available information before conducting the EOGRTS, the registrant includes the extension of Cohort 1B to produce the F2 generation although not requested in ECHA's decision, it should be analysed if a 10-week premating exposure duration for P0 animals is still needed. If the substance shows no potential for delayed steady-state kinetics, a duration of at least two weeks should be applied. The 10-week premating exposure duration is covered by the exposure of the F1/P1 generation of Cohort 1B, which is extended to produce the F2 generation.

4.5 Age of animals at mating

According to OECD TG 443, the age of the P0 animals should be similar and approximately 90 days (13 weeks) at mating. P1 males and females are cohabited beginning on or after PND 90, but not exceeding PND 120. However, considering a premating exposure duration of 10 weeks for the P0 generation, ECHA Guidance[4] states that "*the exposure can be started when the animals are around 5 weeks old and mate them around 15 weeks of age.*"

Spermatogenesis starts in 5-6-week old male rats. Therefore, if the 10-week premating exposure duration starts too early for P0, there is no exposure to the test item for the full cycle of spermatogenesis. However, also P0 animals younger than 5 weeks were used at the initiation of the study in around 10 % of the studies. In six studies, the animals were less than 5 weeks of age when mated.

Also F1 animals should be exposed for 12 weeks from weaning, to have a comparable exposure from 5 weeks of age to mating to cover the full cycle.

To allow a meaningful assessment of the effects on sexual function and fertility, the P0 animals should not be younger than five weeks when premating exposure starts. This is required to guarantee that animals are exposed to the test item for a full cycle of spermatogenesis and folliculogenesis. The P0 animals should not be too old either (not exceeding 120 days) when mating starts.

If Cohort 1B is extended to produce the F2 generation, the P1 animals should be mated from PND 105 (not exceeding PND 120) to also guarantee that Cohort 1B animals are exposed to the test item for a full cycle of spermatogenesis and folliculogenesis.

4.6 Requested study design

ECHA's decisions specify key elements of the study design including the extension of Cohort 1B, and the DNT and DIT cohorts if triggered. All studies but one met or exceeded the requested study design. The single study not meeting the requirement omitted the mandatory Cohort 1B. In 12 studies, the design was expanded beyond what was requested in the respective decisions.

Eight studies additionally included the extension of Cohort 1B, two studies included the DNT cohorts, and two studies the DIT cohort. The following justifications for extending Cohort 1B to produce the F2 generation were given: (*i*) To confirm effects on sexual function/fertility observed in P0, (*ii*) to confirm developmental effects observed in F1, and (*iii*) to evaluate equivocal developmental effects observed in F1. In some cases, however, no justification for extending Cohort 1B was presented, and no justifications were provided for adding the DNT and DIT cohorts.

Cohort 1B is a mandatory part of the EOGRTS and cannot be omitted.

The study design, including any added expansions, must be fully justified and documented. Further detailed guidance on the study design and triggers is provided in ECHA Guidance R.7a on IRs & CSA, Section R.7.6.

If registrants decide to expand the EOGRTS by including the extensions of Cohort 1B to produce the F2 generation before initiating the study, they should also analyse if the premating exposure duration can be shortened to at least two weeks as already explained.

4.7 Mandatory investigations not conducted or conducted with deviations

This section addresses the splenic subpopulation analysis, as well as investigations on sexual maturation and thyroid hormones.

Splenic subpopulation analysis: In two studies, the mandatory splenic lymphocyte subpopulation analysis in Cohort 1A was omitted. If this investigation is missing, information on the potential (developmental) immunotoxicity of the substance is limited, especially in cases where the DIT Cohort 3 is not requested.

Investigations on sexual maturation: According to OECD GD 151, sufficient animals (three/sex/litter/dose) should be maintained until sexual maturation and a combined statistical analysis should be performed for this investigation. However, it was noted that not all laboratories followed these specifications. The issues identified were: (*i*) In around 20 % of studies, the data from different cohorts were not combined for analysis of sexual maturation; and (*ii*) around half of the studies did not evaluate the required 60 pups/sex/dose. Furthermore, it seemed that in some cases the statistical unit for analysing sexual maturation was not the litter. It is nessesary to specify whether the litter effect was taken into account in the statistical analyses and if a nested design was used. If the statistical analysis uses the individual pups as the statistical unit instead of litter, this will lead to the use of an excessively high N value, and could lead to wrong statistical conclusions.

Thyroid hormones: OECD TG 443 and GD 151 specify that TSH and T4 levels should be measured in 10 animals/sex/group in P0 and F1A animals at termination or at a pre-termination bleed, and F1 surplus and F2 weanlings at termination, and optionally, F1 surplus pups at PND 4. In some studies, however, the thyroid hormone measurements were not conducted in all life stages and/or sexes as required. The thyroid system can be differentially targeted at different life stages and, therefore, a lack of effects in adults should not be used to justify exclusion of thyroid system hormone measurements at other life stages. Furthermore, both sexes should be evaluated separately for the thyroid hormone levels as there may be sex-specific responses to the test substance that cannot be identified if only one sex is evaluated or if the samples are pooled for the analysis.

The splenic subpopulation analysis in F1A is a mandatory investigation.

60 pups/sex/dose must be evaluated for sexual maturation. The litter remains the statistical unit for analysing this data.

TSH and T4 levels must be measured for P0 and F1A at termination and for F1 and F2 surplus weanlings, optionally for F1 PND 4 surplus pups. Both sexes must be evaluated separately.

4.8 Investigations in the F2 generation (mandatory when F2 triggered)

If the extension of Cohort 1B is triggered, ECHA's decisions normally request to keep the F2 animals until weaning.

Annex table 1.2 of OECD GD 151 clarifies that endpoints and examinations required in F1 litters are identical to F2 up to weaning. Special attention should also be paid to any target organs identified and the analysis of TSH and T4 to adequately address possible concerns in F2. Only conducting identical investigations in F1 and F2 enables an adequate comparative analysis of the first and second filial generation. Therefore, F1 and F2 must be subjected to identical investigations if an extension of Cohort 1B is triggered. This may be of significant value for setting the NOAEL for the offspring (and the lowest NOAEL for the study) and may strengthen the conclusion for the presence and severity of hazard and the need for hazard classification.

Investigations in the F2 generation must be identical to those of the F1 generation up to weaning.

4.9 Conducting additional investigations

To follow up a concern identified before or during the study, the test laboratories may add additional investigations. In this respect, OECD GD 151, paragraph 55 specifies that "*there are many possible endpoints that could be included (see paragraph 56), but care should be taken when considering these so that they do not compromise the standard endpoints described in TG 443 (see Annex 1 of this document).*" If additional investigations are included in the study without being requested in the decision, the reason for those should be clearly stated.

Additional investigations not described in OECD TG 443 were conducted in multiple studies including e.g. blood prothrombin time blood test in F1, or other additional clinical chemistry and haematology parameters including steroid hormone and luteinising hormone measurements, histopathology of lymph nodes and bone marrow as well as enumeration of ovarian follicles and corpora lutea in the parental generation (in addition to F1 generation), investigation of nipple retention at multiple time points instead of only one on PND 12/13 as specified in the TG, investigation of testicular spermatid counts, as well as mechanistic information such as liver and plasma choline concentration as well as additional histopathological staining of tissues.

Additional mechanistic information can be useful for elucidation of the mechanistic background of observed effects. In addition, for purposes of comparison, it is reasonable to compare histopathological findings in target organs from different generations. Testicular sperm counts are valuable, especially where changes are detected in the cauda epididymal sperm counts as this can provide further clarification on the target organs. It is recommended that the testicular sperm counts are performed in the EOGRTS, for the same animals as cauda epididymal sperm counts.

For nipple retention, evaluators noted that due to the sensitivity of nipple retention on day 12 as a marker of anti-androgenic activity[22] of the substances, the analysis of nipple retention at later time points should not be used as a surrogate measurement instead of more tedious counting of nipples/areolae on PND 12 (e.g. quantification of the number of nipples for individual pups instead of incidence of male pups with retained nipples). On the other hand, information on the permanency of nipple/areolae retention provides relevant information for hazard identification as while nipple retention in general provides information on the potential anti-androgen action of the substance, permanent nipple retention in males is considered a malformation.

The testicular sperm counts should be performed for the same animals as cauda epididymal sperm counts.

The counting of nipples/areolae should be done at least on PND 12.

4.10 Lactational transfer and direct dosing

OECD TG 443 does not require toxicokinetic data from previously conducted studies but highlights its usefulness in the planning of the study especially regarding exposure-based information on maternal blood and foetal/pup blood and milk levels. Modifications to the study design should be considered when excretion of the test chemical in milk is poor and where there is a lack of evidence for a continuous exposure of the offspring. In these cases, direct dosing of pups during the lactation period should be considered (OECD TG 443, paragraph 26). However, this procedure is challenging and, therefore, there is a need for trained laboratory technicians to avoid unnecessary suffering and death.

If pups do not receive the substance in milk or by direct dosing, there is a gap in exposure during a potentially critical window of development, from birth until the pup starts to eat for itself (in dietary studies) or when direct dosing commences (gavage studies) typically at weaning. It is highlighted that a significant part of brain development happens in neonatal rats after parturition whereas the human brain develops largely in utero. Therefore, direct dosing of pups should be considered if the test item is not present in breast milk, i.e. if the pups are not exposed to the test item for significant periods during lactation. OECD GD 151, paragraph 24 indicates how lactational transfer can be investigated.

Lactational transfer to milk was evaluated in only four studies. In one study, the test item was not detected in breast milk and, therefore, pups were directly dosed by gavage from PND 7 to 20. There was a high mortality in dose groups during the time when pups were directly dosed with evidence pointing to gavage errors. In another report, the test item secretion in the milk was considered likely based on the high lipophilicity of the test substance and the high nominal dose levels. However, most studies did not provide any consideration for the lactational transfer.

[22] Schwartz et al. 2021: On the use and interpretation of areola/nipple retention as a biomarker for anti-androgenic effects in rat toxicity studies. Front. Toxicol., 27 October 2021: https://doi.org/10.3389/ftox.2021.730752

If existing information indicates that there may be a concern on the effects of the substance on the developing offspring from parturition until weaning, ensure that there is no gap in exposure during the potentially critical window of development.

Direct dosing of pups by gavage requires special attention and should always be carried out by adequately trained laboratory technicians to avoid unnecessary suffering and death.

5. Recommendations on reporting and methodologies

In this chapter, issues related to methodologies used in the EOGRTS are presented and discussed with recommendations given for improvements if considered relevant. The issues discussed here are based on a selection covered by the questionnaire (see Annex) used for evaluating the studies in the project. They do not cover all various methodologies and approaches used in the EOGRTS.

5.1 Reporting of methodologies

For a proper analysis and interpretation of results, it is necessary that the applied methodologies are adequately described. In brief, the method section should provide sufficient documentation on how the investigations were conducted, the methods used and reasons for choosing the methods. The full study reports should contain historical and positive control data to demonstrate that the laboratories have technical proficiency to generate reliable and reproducible data and that the methodology is sensitive enough to detect changes in the parameters evaluated. When commercially available kits or devices are used, these should be identified, as well as the possible computational software used. Differences in the methodology for measuring the parameters may cause inconsistencies in results between different studies, rather than inconsistencies in the results themselves.

The methodological descriptions in the reports are very variable. Some study reports describe well the details of the methodology and provide information on the validation of the methodology in the test laboratory as well as historical control information. However, in many cases, the reporting is very limited, and deficiencies in reporting have been identified which may affect interpretation and acceptance of results. Some findings are summarised here below, and further details are provided under the specific sections below.

Registrants and test laboratories should describe the methodology in sufficient detail.

The information should allow an assessment if the laboratory has sufficient experience in conducting the assays. Therefore, existing historical and positive control data should be provided.

5.2 General reporting in IUCLID

Reporting in the IUCLID study record (robust study summary) is less detailed compared to the full study report, which contains thousands of pages with group and individual data. The IUCLID study record may include some critical data summarised in tabular form but including all the summary tables and individual data from the full study report is not feasible. However, registrants can always attach the whole study report or data tables to the IUCLID study record which are not visible on the dissemination page and, therefore, are available only for authorities to support the evaluation of the provided study.

Understandably, an independent assessment of the results is not possible without numerical data and statistical details. Also, a description of the methodologies requires sufficient information to understand how investigations were performed. Generally, the mean value with

standard deviation and number of cases are a sufficient level of information, but occasionally individual data is needed (e.g. to identify outliers, to analyse if specific effects occur in a few litters only or are distributed over many litters).

During the evaluation, information in the IUCLID study records was compared with corresponding information in the full study reports regarding effect information and description of methodologies. The aim was to assess if the IUCLID study records provide sufficient information for independent assessment and drawing conclusions.

Generally, the IUCLID study record was in line with the information in the full study report. However, certain findings were indicated in the full study reports but not described in the IUCLID study records.

In most cases, the study record in IUCLID alone was sufficient, but in 17 cases (31 %) the IUCLID study record was not sufficient to draw conclusions. The main deficiencies were:

- A lack of or limited numerical data available in the IUCLID study record to allow independent evaluation of information.
- The reporting in the IUCLID study record was insufficient to give a clear picture of the observed effect, and it may not have been possible to conclude on the adversity and regulatory relevance of all the effects based on the information in IUCLID.
- Some statistically significant findings were not reported in the IUCLID study record.

As an example, numerical toxicological data is not provided for all results in IUCLID, even if “statistically significant effects” are reported. It is often only stated “no treatment-related effects”. Whether an effect is considered treatment-related or not, is influenced by the interpretation of data and may differ between evaluators. Therefore, the numerical data (e.g. mean ± SD) should be provided also in the robust study summary to allow independent evaluation.

In addition, in some cases the IUCLID study record states “no statistically significant effects”. While not statistically significant, some changes may still be considered biologically significant/ relevant, in particular if they follow a dose-response relationship or are outside the historical background. A lack of statistical significance does not exclude biological significance/relevance. Therefore, in cases where there are clear differences in parameters between the treatment groups, the numerical information should be provided to allow confirmation on the lack of biological relevance of the findings.

Based on the identified challenges in performing successful evaluation, it is recommended to attach the full study report in IUCLID, especially for complex studies, such as the EOGRTS but also for all other studies.

The IUCLID study record should not only be a qualitative description of the findings but always mention quantitative results (incidence, magnitude) such as means, standard deviations, coefficients of variation, confidence intervals, etc. For developmental effects it is also important to state litter incidences (if observed effects are confined to one/few litters or occur in several/many/all litters).

5.3 Counting of corpora lutea, primordial and small follicles

In the EOGRTS, ovarian histopathology and enumeration of ovarian follicles and corpora lutea (CL) in F1 adults are the only measures of fertility in females exposed in utero if Cohort 1B is not extended to produce the F2 generation.

The main deficiencies identified in the enumeration of ovarian follicles and CL included:

- Lack of information on laboratory proficiency.
- High variation.
- No CL quantification, quantification of only primordial follicles.
- Lack of confirmation of equivocal findings using second ovary/Cohort 1B.

OECD TG 443, paragraph 73 and OECD GD 151 paragraphs 73 and 74 provide general guidance on methodological aspects.

The information on the laboratory proficiency to conduct enumeration of follicles and the expected sensitivity of the methodology to detect treatment-related effects including historical and positive control data are critical. However, none of the studies included information on the laboratory proficiency to conduct enumeration of follicles and the expected sensitivity of the methodology.

The lack of information on the laboratory proficiency as well as limited or absent methodological descriptions for counting CL, primordial and small follicles hampered the evaluation of the provided results. The methodological descriptions were generally very limited and all details were absent in 18 cases.

The variation in the follicle counts were very high (coefficient of variations up to 100 % in some cases). Due to the limited information on the methodology used and the lack of any information on the laboratory proficiency in the methodology, it is not possible to conclude if this variation is due to biological variation, inappropriate methodology or limited skills in processes involved in the ovarian follicle counting (such as preparation of sections, light microscopy). Laboratories should aim to improve the reporting details on the methodology and training for the laboratory personnel to reduce the potential methodological and observer error as much as possible.

OECD TG 443 specifies that "*a histopathological examination should be aimed at detecting a quantitative evaluation of primordial and small growing follicles, as well as corpora lutea, in F1 females".* This means that quantitative evaluation of follicles as well as CL is expected. However, numerical data for the CL in F1 ovaries was not provided in 18 studies. In six cases, only primordial follicles instead of primordial and small growing follicles were quantified, whereas in an additional five cases it is not clear if primordial and small growing follicles were investigated.

In four cases, the follicle and/or CL counts were non-significantly changed and the numbers in the high-dose group were less than 85 % of the control. OECD GD 151 recommends for the second ovary to be processed if these conditions occur. However, only in one of these cases, did the laboratory evaluate the number of follicles in both ovaries. However, in this case, the number of sections for each ovary was very low (five sections/ovary). In addition, Cohort 1B was not used to further investigate the equivocal findings.

Quantitative evaluation of both primordial and small follicles as well as CL should be performed.

The methodology used for counting follicles should be clearly explained. The description of the methodology should allow evaluation of how the sections from the ovaries were selected for evaluation, and if the section sample size was statistically appropriate for the evaluation procedure used.

Data tables and summaries should include sufficient information to render the findings meaningful (definitions on units of measurement such as average number per section or total number per section evaluated including number of sections evaluated).

Historical and positive control data providing information on the proficiency on the methodology to detect treatment related effects should be provided.

The recommendations in OECD GD 151 to clarify equivocal findings should be followed.

5.4 Oestrous cycle

Information on oestrous cycling can be a useful indicator of the normality of reproductive neuroendocrine and ovarian function in non-pregnant females. The oestrous cycle stage at the termination also allows interpretation of hormonal, histologic, and morphologic measurements relative to stage of the cycle. OECD TG 443, paragraph 80 requires that results from oestrous cycle investigations should be reported as "*number of P and F1 females with normal or abnormal oestrous cycle and cycle duration*". Information on the cycle phase needs to be determined also at termination to enhance assessment of endocrine sensitive organs.

It is also important to consider the regularity of the oestrous cycles before starting the reproductive toxicity studies as high proportion of non-cycling females or females with irregular cycles may cause difficulties in interpretation of study findings, not only on the oestrous cyclicity but also parameters such as time to mating and fertility incidence. OECD TG 443, paragraph 15 also specifies the need for assessing the oestrous cycle before the start of treatment.

In most cases, the information on the cycle duration and/or normal and abnormal oestrous cycle were reported. However, in around 25 % of the studies, no information was provided on the regularity of the individual oestrous cycles. Evaluators also emphasised that the criteria used for the categorisation such as regular, irregular, and acyclic should be included in the report and the incidences for regular, irregular, and acyclic animals per group should be provided.

Laboratories evaluated the mean cycle length for each animal and summarised the mean cycle length for each experimental group. In some cases, the reports also contained information on the mean duration of each cycle stage. Lengthening of the cycle may be a result of increased duration of specific stage of the oestrous cycle such as oestrus or dioestrus. Knowing the affected phase can provide direction for further investigation.

It is important that the mean cycle lengths from all females in the group are used for the calculation of the mean cycle duration in the treatment groups, and not only the females with regular cycles. Evaluators noted that in some cases only the cycle duration from the regular cyclers seem to have been used. This can significantly impair the conclusions if only the summary data is reported.

The mean cycle lengths from all females in the group should be used for the calculation of the mean cycle duration in the treatment groups, and not only the females with regular cycles.

Reporting should include information on incidences of females with normal or abnormal oestrous cyclicity such as regular, irregular, and acyclic. In addition, mean number of days in each stage (oestrus, metoestrus, dioestrus and prooestrus) or the % of cycle at each stage should be reported.

The criteria used for the categorisation of regular, irregular, and acyclic should always be included in the study report.

5.5 Sexual maturation

OECD TG 443, paragraph 47 explains that the F1 animals are evaluated daily for vaginal patency or balano-preputial separation and compared to physical development, i.e. age and body weight. Further information on the onset of puberty is provided in OECD GD 43 which describes that vaginal opening and balano-preputial separation occur in rats around PND 30-35 and 40-45, respectively. It also explains that training of staff is necessary to ensure similar scoring of animals.

The median age at vaginal opening (median of study means) for control Wistar Han rats was around 31.6 days with the mean ranging from 29 to 42 days, while the median for control Sprague Dawley rats was 34 days ranging from 33 to 36 days. The variation in Wistar Han rats

was much greater compared to Sprague Dawley. This was also observed for the age at achieving balano-preputial separation: the median age of control Wistar Han rats was around 42 days ranging from 32 to 51 days while in control Sprague Dawley rats the median age was 44.5 days ranging from 42 to 47 days.[23]

In most cases, the vaginal opening and balano-preputial separation were observed starting from PND 25-28 and 35-38, respectively, depending on the laboratory. In one study, the evaluation of sexual maturation was initiated only when some females had already achieved the vaginal opening or the day at vaginal opening, and in another study when all males had already achieved the balano-preputial separation. This raises uncertainty on whether some females in the study would have met the criteria for vaginal opening already before the start of observation.

In another study, the ages at vaginal opening and balano-preputial separation were clearly outside the expected range and pattern because the mean age for vaginal opening was PND 40.2 and for balano-preputial separation PND 32.3. No historical control data was provided for the laboratory to confirm the validity of the measurements.

In some cases, the body weight on the day of meeting the criteria for vaginal patency or balano-preputial separation was not reported. This impairs the interpretation of findings as the sexual maturation in the context of general (delayed) development of the animal as the body weight is used as an indicator of physical development. However, in most cases, the mean age at sexual maturation and the corresponding mean body weights when meeting the criteria were reported in the summary tables. Some studies also presented the data as a graph showing the percentage of evaluated animals which had reached the criteria at certain days (days on the x-axis and % of animals that had met the criteria on the y-axis). In addition, in some reports the percentage data was presented in a tabulated form. Evaluators considered that this 'survival' type (i.e. Kaplan-Meier curves) of presentation was very informative for the evaluation and could better show the slight shifts between the groups.

Investigations on sexual maturation must be initiated early enough to not miss individuals that achieve this landmark early.

When the ages at vaginal opening and/or balano-preputial separation are clearly outside the expected range/pattern, the laboratories should provide an explanation. The actual age (postnatal day) should be reported.

High variation in the age reaching the criteria and unexpected results highlight the need for historical control data from the laboratory.

Survival type presentation of sexual maturation (graph showing the percentage of evaluated animals which had reached the criteria at certain days) is informative and can show slight shifts in timing between the groups.

5.6 Grouping of organs for weighing

According to OECD TG 443, some reproductive organs and accessory organs can be weighed together, notably uterus with oviducts and cervix, and seminal vesicles with coagulating glands and their fluids.

As an example, it was noted that prostate and seminal vesicles with coagulating glands were weighed as a whole, instead of prostate separately from the seminal vesicle and the coagulating gland. In addition, the uterus was weighed with cervix, and the ovaries were weighed with

[23] This analysis is based on the first 36 cases (21 Wistar Han and 15 Sprague Dawley) that were evaluated for this project.

oviducts, instead of the uterus with oviducts and cervix and ovaries separately.

Combining different organs and tissues that are sensitive to endocrine modes of action can complicate interpretation of potential changes but also conceal potential changes. In addition, different combinations in different studies makes comparing the findings more difficult.

Ovaries are weighed separately. Epididymides should be weighed as a whole and the weight of the cauda epididymis (samples used for cauda epididymal sperm counts) should also be reported separately. In addition, dorsolateral and ventral lobes of the prostate are collected together but dissected after fixation and weighed separately in P0 and F1A, and prostate in Cohort 1B. In the event of a treatment-related effect on total prostate weight, the dorsolateral and ventral segments should be carefully dissected after fixation and weighed separately.

5.7 Sperm parameters

According to OECD TG 443, sperm parameters should be measured in all P0 males and in Cohort 1A males. One epididymis should be used for enumeration of cauda epididymis sperm reserves. In addition, sperm from the cauda epididymis (or vas deferens) is collected for evaluation of sperm motility and morphology. Total cauda epididymal sperm numbers should be reported.

In some cases, the enumeration of sperm reserves was not conducted or were evaluated from whole epididymis (including caput, corpus, and cauda) instead of the cauda epididymis. Total cauda epididymal sperm numbers should be reported together with the cauda epididymis weights. In addition to the total cauda epididymal numbers the sperm numbers can also be reported as sperm number per gram tissue. This allows independent evaluation of the information. Evaluators also highlighted that the sperm parameters should be investigated in all males (all treatment groups).

Cauda epididymis concentrates the sperm and functions as the sperm storage reservoir until the time of ejaculation. The contents of the caput reflect sperm just released from the testis whereas the contents of the cauda reflect sperm that left the testis 4-14 days previously. This is particularly important for interpreting potential treatment-related effects. Therefore, the sperm enumeration should be performed in the cauda epididymis and not the whole epididymis. In addition, the weight of the cauda epididymis, and not only the whole epididymis should be provided.

The sperm enumeration should be performed in the cauda epididymis and not the whole epididymis. Total cauda epididymal sperm numbers should be reported. In addition, the weight of the separate cauda epididymis should be provided.

5.8 Calculation and reporting post-implantation loss

According to OECD TG 443, the results must report the number and percentage of post-implantation loss, live births, and stillbirths. Post-implantation loss is the difference between the number of implantation sites in the uteri and the number of full-term foetuses/pups.

OECD GD 43 specifies that "*Comparison of the number of implantation sites with the number of live and dead foetuses or neonates for each litter provides a means of quantifying post-implantation loss*." The aim of this parameter (i.e. post-implantation loss) is to illustrate mainly in utero deaths and can include some deaths during parturition. This can then be differentiated from the early postnatal death which may be caused not only by developmental toxicity but also due to poor maternal care or lactation-related aspects.

Post-implantation loss can be assessed only if the live foetuses are counted after caesarean

section, before expected parturition, such as in prenatal developmental toxicity studies. In OECD TG 443 studies, and other reproductive toxicity studies where pups are allowed to be born, the number of implantation sites in the uteri can be assessed. However, the number of pups at the time of observation may not adequately reflect the total number of pups (live and dead) born because the pups are allowed to be born and dams may cannibalise any malformed and dead pups. In addition, it may be challenging to identify stillborn pups and pups that died after birth (postnatal loss) if pups have been partly cannibalised. Post-implantation loss is reflecting developmental toxicity, but cannibalism or early postnatal loss may be due to developmental toxicity but also due to maternal reasons and may be considered as sexual function and fertility effects or due to general toxicity.

It is important to try to separate developmental toxicity (post-implantation loss) from adverse effects on sexual function and fertility (postnatal deaths due to poor maternal care), whether or not considered secondary to poor maternal condition.

It was noted that in some EOGRTS reports, post-implantation loss is reported as a difference between implantation sites and the number of live pups on PND 0 and not with total number of offspring, dead and alive. Therefore, in these cases it may be difficult to decide if the observed effect is developmental or due to other factors such as lack of postnatal maternal care. It is also not consistent with the common approach proposed in OECD GD 43.

The post-implantation loss/survival index should be calculated based on the total number of offspring born (dead and alive) with the information available on the incidence of dead offspring.

5.9 Calculating and reporting live births, postnatal loss / viability index

OECD GD 151, paragraph 89 refers inter alia to the following indices *"The major indices usually determined are: male and female mating indices, male and female fertility indices, gestation length, gestation index and survival index. These should be reported in TG 443. Calculation of these indices and discussion on interpretation of reproductive performance can be found in GD 43 (OECD, 2008, paragraph 180)."*

OECD GD 43 includes the indices shown in Table 2.

Table 2: Reproduction indices in OECD GD 43.

Index	*Calculation*
Male Mating Index[24]	No. of males with confirmed mating X 100 Total No. of males cohabited
Male Fertility Index[25]	No. of males impregnating (siring) a female X 100 Total of No. males cohabited
Female Mating Index[26]	No. of sperm positive females X 100 Total No. of females cohabited
Female Fertility Index[27]	No. of pregnant females X 100 No. sperm-positive females
Gestation Index[28]	No. of females with live born pups X 100 No. of pregnant females

[24] Measure of male's ability to mate
[25] Measure of male's ability to produce sperm that can fertilise eggs
[26] Measure of female's ability to mate
[27] Measure of female's ability to become pregnant
[28] Measure of pregnancy that provides at least one live pup

The postnatal viability, such as viability index/postnatal loss on day 4 (before standardisation of litters) should be calculated. The viability index on day 4 is calculated as the number of live offspring on day 4 (before culling) compared to the number of live offspring on the first check (day 0 or day 1).

OECD GD 43 specifies the Survival Index (%) as (No. of live pups (at designated time)/No. of pups born) x 100). In many cases, the laboratories record the Live Birth Index (%) as (Number of offspring on Day 1 after littering / Total number of offspring born) x 100).

For the live birth index (%), useful definitions that are commonly used for describing live births are:

- the number of live pups on PND 0/number of implantations × 100; and
- the number of live pups/numbers of total pups × 100.

These will give an indication of the proportion of live births out of all implantations (also includes post-implantation deaths before birth) or live births out of all deaths. In addition, if live pups were also calculated on PND 1, this would allow the survival index to be calculated for day 1 (% of live pups born surviving to PND 1). However, OECD TG 443 specifies that "*Live pups are counted and weighed individually on PND 0 or PND 1*". Therefore, this index cannot be calculated as such if the live pups are only counted on PND 1.

In some cases, the laboratories use the same terminology, but the calculations use different information. Therefore, in the full study reports and in IUCLID study records, the used reproductive indices should be provided with definitions, usually expressed as calculation formulae.

Explanatory descriptions of the reproductive indices with the brief definitions used for calculations including post-implantation loss and postnatal loss should be provided in the full study report. But they should also be provided in the IUCLID study summary to allow the evaluators to understand what the numbers reflect and help comparing the findings between different studies.

The methodology and criteria used should be consistent between the generations (and dams) to allow interpretation of the results.

If there is extensive cannibalism that seem to affect the parameter, it is proposed that instead of "post-implantation loss" the term "post-implantation and postnatal loss immediately after birth" is used, which better describes the actual situation.

5.10 Anogenital distance (AGD) / Anogenital index (AGI)

According to OECD TG 443, paragraph 46, "*the anogenital distance (AGD) of each pup should be measured on at least one occasion from PND 0 through PND 4. Pup body weight should be collected on the day the AGD is measured and the AGD should be normalized to a measure of pup size, preferably the cube root of body weight [...].*"

OECD GD 151, paragraph 60 further defines: "*TG 443 requires that anogenital distance (AGD) be measured on each pup on at least one time point between PND 0 to PND 4. It is important that all pups are measured on the same postnatal day because the rapid growth of pups will also affect AGD. Further guidance on measurement of AGD is provided in GD 43 (OECD 2008, paragraph 90) and methods of determination of AGD have been recently described by Christiansen et al (2010) and Gray et al (2009).*" OECD GD 43 guides on the measurement and interpretation of the changed anogenital distance.

In the EOGRTS, anogenital distance was measured either on PND 1, 2 or 4. The observed deficiencies in the analyses included large variations in the reported values and a lack of normalised data. With large variation (e.g. due to several persons measuring), relevant changes may not be observed although there may be changes if precision of measurement would be

higher, reducing variation. In addition, while historical control data was included in some reports, no positive control data demonstrating proficiency in the methodology were provided in any of the studies to allow evaluators to determine whether the testing laboratory can detect a treatment-related effect.

While in most cases the anogenital distance was reported as an absolute measured distance (mm) and normalised to the cube root of body weight (anogenital index), in some cases, only the absolute measured distance was provided. This may lead to incorrect interpretation of the findings if there are markedly lower body weights in some groups when compared to controls.

It is important to explain clearly how the anogenital distance has been calculated, and calculations should be consistent between the generations (and dams) to allow results to be interpreted.

Historical and positive control data providing information on the proficiency on the methodology to detect treatment related effects should be given.

5.11 Number of male pups with nipple retention in controls

According to OECD TG 443, paragraph 46, nipples/areolae should be determined at least once from each F1 and F2 male pups: "*The presence of nipples/areolae in male pups should be checked on PND 12 or 13.*" OECD GD 151, paragraph 61 gives more guidance and also refers to OECD GD 43. It should be noted that quantitative count of nipple/areolae is recommended due to insensitivity of qualitative assessment. OECD GD 43, paragraph 91 provides further details on the evaluation as well as on interpretation of the nipple/areolae information.

Most of the studies seem to have an issue with counting/identifying nipples and/or areolae because the incidence of control male pups with nipples/areolae present on PND 12/13 is reported to be zero. As an example, in one study presence of nipples on PND 13 was evaluated for all, approximately 400, male pups with no nipples recorded for any of the males. Similar findings were also seen in most other studies. Evaluators considered that based on the previous experience from many scientific publications this does not seem biologically plausible because some degree of nipple retention is expected even in control males[22]. While historical control data was provided in some reports (especially when nipple retention was detected), no positive control data providing information on the proficiency were provided in any of the studies to allow evaluators to determine whether the testing laboratory can detect a treatment-related effect.

There are no indications that any of the most commonly used rat strains in toxicology testing show differences in background incidence of nipple retention or that any of them show a zero-background incidence.[22] Therefore, if there are indeed test laboratories which consistently report no nipples in any of their male pups, the sensitivity of those test laboratories for detecting nipple retention should be investigated.

Historical and positive control data providing information on the proficiency on the methodology to detect treatment related effects should be given.

5.12 Thyroid hormone measurements

OECD TG 443 indicates from which animals and when the blood samples are to be drawn, however, it does not refer to any methodology for measuring T4 and TSH. OECD GD 151, paragraph 70 highlights the need to examine the assays before accepting them for use and that information should be reported with the results of the assay: "*The performance of the T4 and rodent TSH assays should be examined prior to accepting and examining the thyroid hormone results between treated and control groups for any study. The standard reference curves, level of hormone sensitivity, quality control samples and within- and between-assay coefficients of variation should be within acceptable limits according to the manufacturer's specifications and laboratory SOPs. This information should be reported along with the results of the assay.*"

The EFSA-ECHA Guidance for the identification of endocrine disruptors[29] contains additional recommendations for thyroid hormone analyses including but not limited to anaesthesia, blood sampling, sample storage, assay validation and use of historical control data. In addition, while no specific statistical analysis methodology is recommended, the Guidance specifies that "*High variability should trigger outlier statistics and justification for each excluded data point should be provided*."

Below are some deficiencies on thyroid hormone and TSH measurements that were discussed by evaluators.

The methodologies used for the thyroid hormone and TSH measurements in the evaluated EOGRTS included radioimmunoassays, enzyme-linked immunosorbent assays, and liquid chromatography tandem mass spectrometry. The description of analytical methodologies used for thyroid hormones and TSH measurements should be provided. This should include detailed information on e.g. sampling techniques (time, anaesthesia, pooling samples, storage, processing), as well as the name of the commercial assays if used. In the worst cases, no information was provided on the applied test method. In some cases, the validation of the methodology was described in the full study reports while in other cases only the name of the commercial kit was provided.

Furthermore, validations of analytical methodologies before use are essential. As indicated in OECD GD 151, paragraph 7, the performance of the assays should be validated, and the information of this validation should be reported along with the results of the assay. Further information is also available in the EFSA-ECHA Guidance for the identification of endocrine disruptors[29] and report from the BfR expert hearing on practicability of hormonal measurements[30]. Performance criteria should establish e.g. parallelism of matrix samples to the reference standard, determination of assay sensitivity and measurement range, suitability of the chosen assay for the species used in the study, and precision and use of appropriate positive control compounds. It is also recommended that to demonstrate proficiency for thyroid hormones measurement, a laboratory should be able to show results from a separate study in a comparable test system (assay, rat strain, or blood sampling route) using a positive control substance.

This information is critical for the evaluation of e.g. the validity of the methodology and detection limits (sensitivity) of the method for different life stages. In addition, the information allows evaluation of sensitivity of the test system in relation to the statistical power of the study and factors that could influence on the variability in hormone measurements. Without this information, there is uncertainty in the conclusions related to the thyroid hormone and TSH measurements and this may hinder the assessment of the effects of the substances on the thyroid system, requiring further information to be generated for definite conclusions.

In many studies, the individual variation of the thyroid system hormone and especially the TSH levels were very high. This may be related to the methodological issues or to innate biological variability, or a combination of both. In addition, especially in the offspring, in some studies the TSH values reported were below the quantification limits of the assay. There are no specific performance criteria for the TSH measurement in OECD TG 443. In publications on compiling historical control data for thyroid hormones (T3 and T4)[31], it was found that in most studies, irrespective of animal age, the control coefficients of variation for T4 and T3 were well below the OECD recommended upper boundary of acceptance 25-30 %. However, TSH is an inherently

[29] https://echa.europa.eu/fi/-/guidance-on-identifying-endocrine-disruptors-published
[30] Kucheryavenko et al 2019. Report from the BfR expert hearing on practicability of hormonal measurements: recommendations for experimental design of toxicological studies with integrated hormonal end points. Arch Toxicol. 93(4):1157-1167. doi: 10.1007/s00204-019-02436-3
[31] Li et al. 2019. Practical considerations for developmental thyroid toxicity assessment; What's working, what's not, and how can we do better? Regulatory Toxicology and Pharmacology 106; 111-136: DOI: 10.1016/j.yrtph.2019.04.010

more variable measure. The OECD recommends a coefficient of variation of 35 % as the upper limit based on data derived from adult TG 407.

It is estimated that 10 rats per group has sufficient power to detect a 1.5-fold increase in TSH and a 1.35-fold (35 %) decrease in T3 or T4 concentration in the treated group vs control, assuming that the coefficient of variation in the control and treatment group is about 25 % for T3 or T4 and 35 % for TSH (Wilcoxon-test, two sided, power 75 %, $p < 0.05$, NQUERY software used for the calculation)[31]. In OECD TG 443, thyroid hormones belong under the clinical biochemistry and, therefore, in the parental and F1A animals are measured in 10 animals/sex/group. Evaluators noted that in many EOGRTS the coefficients of variation in control and treatment groups were >60 % for TSH and in some cases also for T4.

The high variability especially in the TSH concentrations identified in most of the EOGRTS emphasises the need for understanding the expected individual variation of the TSH levels as well as the need for further work on the methodological development. Currently, the analyses indicate that at least with the current methodology, the 10 animals/sex/group do not provide sufficient statistical power. This also highlights the need for adequate top-dose level selection as the variability in the data means that the extent of effect needs to be relatively large to be detectable.

Historical control data for thyroid system hormones must be established and lie within a reasonable range that allows biologically significant changes to be detected. These data can be used for documentation of consistency between studies and may serve as a supplemental parameter helping interpretation. Historical control information serves as basis for comparison to conclude if the concurrent control of the EOGRTS is valid or not. However, primary assessment of an effect in treatment groups is always made by comparison with the concurrent study control group.

Regarding sensitivity, it is critical to ascertain that the limit of quantification of any hormone assay is well below control values of the age of animals under study. To allow evaluators to evaluate if it is sufficiently and consistently below control values, information on the limit of quantification for the assay in the specific conditions of the laboratory (and the limit of detection) is needed.

The methodology used for measuring T4 and TSH should be clearly explained.

If the variation in hormone levels in the control groups is higher than indicated in OECD TG 407/408, an explanation should be provided.

To demonstrate proficiency for thyroid hormones measurement, a laboratory should be able to show results from a separate study using a positive control substance. In addition, appropriate historical control data and proof that the limit of quantification of any hormone assay is well below control values of animals under study should be provided.

EUROPEAN CHEMICALS AGENCY
P.O. BOX 400, FI-00121 HELSINKI, FINLAND
ECHA.EUROPA.EU