Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

14-Dec-2015

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

FEASIBILITY STUDY FOR MINOR ENHANCEMENTS OF TG421/422

Series on Testing and Assessment
No. 217

JT03388226

ENV/JM/MONO(2015)24
Unclassified
English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 217

FEASIBILITY STUDY FOR MINOR ENHANCEMENTS OF TG 421/422 (REPRODUCTION/DEVELOPMENTAL TOXICITY SCREENING TEST) / (COMBINED REPEATED DOSE TOXICITY STUDY WITH THE REPRODUCTION/DEVELOPMENTAL TOXICITY SCREENING TEST) WITH ENDOCRINE DISRUPTER-RELEVANT ENDPOINTS

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organisations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Series on Testing and Assessment link

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document includes the Feasibility study for minor enhancements of TG 421/422 (Reproduction / Developmental Toxicity Screening Test) / (Combined Repeated Dose Toxicity Study with the Reproduction / Developmental Toxicity Screening Test) with Endocrine Disrupter-relevant endpoints. The objective of the project, proposed by Denmark in 2013, was to examine the feasibility of inclusion of sensitive endpoints for the detection of chemicals with endocrine disrupting properties in TGs 421/422 and to update the TGs accordingly.

The feasibility report is based on statistical analyses of the proposed additional endpoints – nipple retention, anogenital distance and thyroid. The statistical analyses are available as annexes to the report, which also includes proposals for an update of the version of TGs 421/422 published respectively in 1995 and 1996.

The draft feasibility report was circulated for comments to the WNT in September 2014. Comments received were addressed by the Lead Country, in consultation with the expert group on reproductive toxicity and further thyroid data analysis were conducted early 2015. The feasibility report was updated in February, based on this additional analysis and was approved by the WNT in April 2015, declassified and published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides, and Biotechnology on 10 July 2015.

ACKNOWLEDGEMENTS

The Feasibility study for minor enhancements of TG 421/422 with Endocrine Disrupter-relevant endpoints was prepared by Sofie Christiansen & Ulla Hass (Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark, Denmark). The statistical analysis (annex 1a and 1b) was developed by Martin Scholze, Senior Consultant: Biostatistics (Scholze Consultancy), based on data contribution from the Technical University of Denmark but also from several experts in the OECD expert group on reproductive toxicity, who provided data from their respective laboratories. The OECD expert group on reproductive toxicity also provided support to the lead country during the development of the feasibility report and the revision of the Test Guidelines 421/422, and held several teleconferences between June 2014 and February 2015.

Table of contents

Terms of reference
Aim
Background and expected regulatory need/data requirement that will be met by the proposed outcome of the project
Anogenital distance
- Method
- Data analysis, sensitivity/power,
- Human relevance
- Animal welfare
- Inclusion of AGD in TG 421/422

Nipple retention
- Method
- Data analysis, sensitivity/power
- Human relevance
- Animal welfare
- Inclusion of NR in TG 421/422

Thyroid hormones
- Method
- Data analysis, sensitivity/power
- Human relevance
- Animal welfare
- Inclusion of thyroid hormones in TG 421/422

Abnormalities of external genital organs
- Methods
- Data analysis, sensitivity/power,
- Human relevance
- Animal welfare
- Inclusion of abnormalities of external genital organs in TG 421/422

Overall discussion and conclusions
References

Appendix 1a Power Simulations of Nipple Retention and Anogenital Distance of Rodents exposed to Endocrine Disrupting Chemicals
Appendix 1b Power Simulations of T4 hormone levels of Rodents exposed to Endocrine Disrupting Chemicals
Appendix 2 Study report from DTU Food on malformations of the external genitalia in young male rat offspring.

Terms of reference

1. This draft report has initially been prepared by the National Food Institute, Technical University of Denmark. The report gives input for discussions in the OECD working group of experts involved in the project Feasibility study for minor enhancements of TG 421/422 (Reproduction/Developmental Toxicity Screening Test) /(Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test) with ED-relevant endpoints. Subsequently, the draft report has been revised based on the discussions in this working group.

AIM

2. The aim of this project is to do a feasibility study for minor enhancements of TG 421/422 (Reproduction/Developmental Toxicity Screening Test) /(Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test) with ED-relevant endpoints. This review addresses scientific and technical concerns regarding inclusion of additional ED related endpoints in TGs 421/422. The endpoints considered include anogenital distance (AGD), Nipple Retention (NR), thyroid hormones and malformations of external reproductive organs in male offspring. For these endpoints, the scientific and technical questions considered include:

- Are standardized methods available?
- Is the sensitivity sufficient with the number of litters per group?
- Are the endpoints of relevance for humans?
- Are there animal welfare concerns?
- Is the enhancement possible without changes or with only minor changes in study design?

BACKGROUND AND EXPECTED REGULATORY NEED/DATA REQUIREMENT THAT WILL BE MET BY THE PROPOSED OUTCOME OF THE PROJECT

3. The TGs 421/422 (Reproduction/Developmental Toxicity Screening Test) /(Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test) provides information on adverse effects on development and reproduction including effects on endocrine organs and is used in various regulatory frameworks (such as REACH) to generate information for risk assessment of chemicals. In GD 150 (Guidance Document on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption) it is written: *"The reproduction/developmental screening tests OECD TG 421 and 422 are included in Level 4 as supplemental tests because they give limited but useful information on interaction with endocrine systems. EDs may be detected by effects on reproduction (gestation, gestation length, dystocia, implantation losses), genital*

malformations in offspring, marked feminized AGD in males, changes in histopathology of sex organs or effects on the thyroid gland" (OECD 2012).

4. However, it is recognized that these *in vivo* screens need to be updated in relation to inclusion of some sensitive effect endpoints relevant for Endocrine Disruption. In the revised OECD Conceptual Framework (CF) from OECD the reproduction/developmental screening tests TGs 421 and 422 are included in Level 4 "if enhanced" as supplemental tests because they provide limited but useful information on interaction with endocrine systems (OECD 2012).

5. DK has undertaken the examination of available existing data and peer review scientific relevant papers to make a proposal to the Validation Management Group on mammalian testing (VMG-mammalian) /Expert group (EG) on reproductive toxicity, on whether or not it is relevant to include these ED related endpoints in a proposal for revision of OECD TG 421 and 422.

6. It will also be considered whether certain slight adaptions of the test designs of these test guidelines may be warranted to include for consideration other ED related endpoints if such are being suggested by the EG/VMG-mammalian for this project.

7. The results of the project may contribute to an improved sensitivity for identification of developmental toxicants in mammalian species at an early stage in the regulatory testing schemes for industrial chemicals (e.g. REACH) as information from TGs 421/422 are already required in such regulatory testing schemes.

8. If these endpoints are implemented in these TGs it will enhance the international harmonization of hazard assessment with regard to developmental toxicity effects (OECD 2012).

9. An important point is that the ability for detection of EDs can be enhanced without increasing the number of experimental animals used.

10. Assessment of AGD and NR are mandatory in TG 443 and could probably easily be included in the TG 421/422. For the examination of NR it is, however, needed to extend the study period in 421/422 from PND 4 to PND 12 or13 to examine this endpoint at the optimal time period.

11. The OECD TG 407 (Repeated dose 28- day oral toxicity study in rodents) has been updated in 2008. The assay has been validated for some endocrine endpoints but the sensitivity of the assay is not sufficient to identify all EATS-mediated EDs. The validation of the assay (OECD, 2006) showed that it identified strong and moderate EDs acting through the ER and AR; and EDs weakly and strongly affecting thyroid function. It was relatively insensitive to weak EDs acting through the ER and AR. This assay also has some optional endpoints such as uterine and ovary weight, changes in vaginal smears, histopathologic changes in mammary gland histopathology as well as serum T3, T4, TSH as well as thyroid weight which can be examined if there is additional concern.

12. The new extended one-generation reproductive toxicity study (EOGRTS) (OECD TG 443) includes more endpoints sensitive to endocrine disruption than OECD TG 416 and, as it also uses reduced animal numbers if conducted without F2, it is expected that it will often replace OECD TG 416 for mammalian reproductive toxicity testing (GD 150). Endpoints sensitive to endocrine disruption, not specified in OECD TG 416, include anogenital distance at birth, areola/nipple retention, measurement of thyroid hormones and TSH levels. Effects on the developing nervous and immune systems are also assessed by the DNT and DIT cohorts. These systems may also be sensitive to endocrine influences. This test is also expected to have greater sensitivity than OECD TG 416 as it requires an increased number of pups to be examined. In summary, the new EOGRT study (OECD TG 443) is preferable for detecting endocrine disruption because it provides an evaluation of a number of endocrine endpoints in the juvenile and adult F1, which are not included in the 2-generation study (OECD TG 416) adopted in 2001.

13. This review also focuses on genital malformation. In TG 443 all selected F1 animals are evaluated around sexual maturity and notes are taken for any abnormalities of genital organs, such as persistent vaginal thread, hypospadias or cleft penis. In the current TG 422 it is noted that each litter should be examined as soon as possible after delivery to establish the number and sex of pups, stillbirths, live births, runts (pups that are significantly smaller than corresponding control pups), and the presence of gross abnormalities.

14. The power of the update endpoints in TG 421/422 with around 8 litters per group compared with the power for similar endpoints in OECD TG 443 with around 20 litters per group is important to consider. This has been done by conducting statistical analyses of existing data (cf. Appendix 1a).

15. TG 407 (Repeated Dose 28-Day Oral Toxicity Study in Rodents) was enhanced in 2008 with regard to inclusion of some ED relevant endpoints. However, it seems even more relevant to also include some ED relevant endpoints in TG 421/422 where the exposure periods cover some of the sensitive periods during development (pre- or early postnatal periods).

16. AGD and NR have the last decades been shown to be sensitive and non-invasive endpoints, when investigating effects of anti-androgenic compounds administered during the critical periods of prenatal development (Clark et al. 1990, Gray et al. 1999, McIntyre et al. 2000, Mylchreest et al. 1999, Hass et al. 2007).

17. Animal studies indicate that both AGD and NR are sensitive markers for increased risk of malformations of the external reproductive organs (Christiansen et al. 2008, Bowman et al., 2003, McIntyre et al., 2002; Welsh et al 2008). Moreover, AGD and NR examinations have been included in the new TG 443 (Extended One-Generation Reproductive Toxicity Study) and in both GD 43 and GD 151 it is stated that AGD can be used for NOAEL setting (OECD 2008, OECD 2013).

Anogenital distance (AGD)

Method

18. New-born male rats have no scrotum, and the external genitalia are undeveloped, and only a genital tubercle is apparent for both sexes. The AGD is the distance from the anus to the insertion of this tubercle, the developing genital bud. The AGD is androgen dependent, and studies show that the AGD is normally about twice as long in male as in female rats. Similarly, in new-born humans the AGD measure was about two-fold greater in males than in females (Salazar-Martinez et al. 2004).

19. The method for assessing AGD is already described in para 45 in TG 443, i.e.:

> *45. 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 (12).*

20. Some further guidance is given in GD 34, i.e.:

> *165. AGD may be influenced by the size of the animal and this should be taken into account when evaluating the data. The size or length of the pups is normally not measured (sometimes crown-rump), but body weights are measured. In some cases, the anogenital index, i.e., AGD divided by body weight, is used. However, body weights of pups may be quite variable leading to a large variation in the anogenital index. This could mask eventual effects on AGD and is therefore not recommended. Instead, the size of the animals should be accounted for by including a covariant. Body weight can be used, but this parameter is in three dimensions, while AGD is in one dimension. Consequently, the optimal covariate seems to be the cube root of the body weight (Clark, 1999). A statistically significant change in AGD that cannot be explained by the size of the animal indicates effects of the exposure and should be used for setting the NOAEL.*

21. In GD 150 it is written: *"For example, feminized AGD in male offspring (observed in OECD TG 416 and possibly in OECD TG 421/422) may be considered as conclusive evidence of an endocrine disrupting effect".*

22. Thus changes in AGD in the OECD 421/422 screening studies can be used for setting an NOAEL. However, if the result is not reproducible in larger, more definitive studies (e.g., OECD 443), the results may be overridden, depending on a case by case evaluation including e.g. the dose levels used in the two types of studies.

Data analysis, sensitivity/power

23. Power Simulations of Nipple Retention and Anogenital Distance of Rodents have been made and are referred to in appendix 1a. These power simulations can be used to calculate the minimum sample size required, in order to likely detect an effect of a given size on the endpoint (e.g. AGD). Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size.

24. For continuous endpoints like AGD the statistical power for detecting significant effects depends on the group size, and on the coefficient of variation in the control group. The effect size needed for having at least 80% probability for detecting significant effects ($p < 0.05$) of a given size on AGD is described in details in Appendix 1a.

25. The results based on both the Copenhagen studies (data from Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark) and the non-Copenhagen studies (Data from other labs) shows that the detection of a 5% reduction in male AGD can be ensured only with at least 16 litters per group. The likelihood for detection of a 10% reduction in male AGD is very high with 8 litters per group.

Human relevance

26. In rats, both AGD and nipple retention has been shown to be highly predictive of adverse effects of the male reproductive system including increased incidence of hypospadias, testosterone decrease and altered reproductive organ weight changes (Bowman et al. 2003, Christiansen et al. 2008, Macleod et al. 2010, van den Driesche et al. 2011, Welsh et al. 2008).

27. Anogenital distance has been shown to be associated with adverse health effects in humans (Bornehag et al 2014). Recent studies reported that male infants and boys with hypospadias or undescended testis had reduced AGD (Hsieh et al. 2012; Hsieh et al. 2008; Jain and Singal 2013; Thankamony et al. 2013). Moreover, a shorter AGD in adult men has been related to decreased fertility (Eisenberg et al. 2011), impaired semen quality (Mendiola et al. 2011) and decreased serum testosterone levels (Eisenberg et al. 2012). Shortened AGD has also been suggested as a biomarker of testicular dysgenesis syndrome (Sharpe 2005).

28. AGD is included as an endpoint in OECD TG 443 and can therefore be considered as an endpoint evaluated to be of human relevance. Moreover ECHA have in a novel evaluation stated that: "The findings in AGD, nipple retention and foetal T, suggest an anti-androgenic mode of action (androgen deficiency) and may be considered as relevant findings and predictors of potential adverse effect during human development." (ECHA 2013). In addition, the OECD GD 43[1] and GD 151 states "A statistically significant change in AGD that cannot be explained by the size of the animal indicates effects of the exposure and should be considered in setting the NOAEL" (OECD 2008; OECD 2013). As the NOAEL can be used as the point of departure for setting safe exposure levels for humans this further supports that effects on AGD are of human relevance. Last, but not least the observations of similar effects in experimental animals and in humans support that effects on AGD in experimental animals are relevant for humans.

[1] *OECD GD 43 (GD on Mammalian Reproductive Toxicity Testing and Assessment; OECD 2008c) states, "A statistically significant change in AGD that cannot be explained by the size of the animal indicates effects of the exposure and should be used for setting the NOAEL".*

Animal welfare

29. An important point to remember is that the ability for detection of EDs can for these tests (OECD TG 421/422) be enhanced without increasing the number of experimental animals used.

30. Assessment of AGD requires slightly more handling of the new-borns. This assessment can be done very gently and is therefore not expected to lead to any animal welfare concerns.

Inclusion of AGD in TG 421/422

31. There are standardized OECD test methods for assessing AGD and the sensitivity analysis shows that relevant data can be obtained with the number of litters per group in the TGs 421/422. Also, AGD is an endpoint of high human relevance and there are no concerns for animal welfare related to the assessment of this endpoint. AGD is normally measured at birth (e.g. PD 1-4) and therefore this endpoint can be included in TGs 421/422 without any modification of the overall test design.

32. This all supports that assessment of AGD can be included in TGs 421/422 (OECD 2015a) (OECD 2015b).

Nipple retention

33. Mammary gland development begins similarly in male and female rats; however, the further development of the nipple is sexually dimorphic (Kratochwil 1971). Female rats have nipples, whereas male rats possess only rudimentary mammary glands but no nipples. This is because locally produced dihydrotestosterone (DHT) causes regression or apoptosis of the nipple anlagen in male rats (Imperato-McGinley et al. 1985; Imperato-McGinley et al. 1986). However, foetal exposure to anti-androgens can block this process, and the male offspring displays nipples similarly to their female littermates. Therefore, the retention of nipples in male rat pups is an indicator of impaired androgen action during the development.

34. Assessment of nipple retention (NR) on postnatal day 12 or 13 is included in TG 443. As TG 421/422 stops on postnatal day 4, we have studied the possibility for assessing NR at an earlier time points, e.g. at birth or on postnatal day 4. This does not appear possible and thus inclusion of NR in these guidelines would require a 10 days extension of the study period, e.g. until postnatal day 13.

Method

35. Generally, nipples/areolas is defined as a dark focal area (with or without a nipple bud) located where nipples are normally present in female offspring (Hass et al. 2007). The method for assessing NR is already described in para. 45 in TG 443, i.e.:

The presence of nipples/areolae in male pups should be checked on PND 12 or 13.

36. Some further guidance is given in GD 151, i.e.:

Para. 61. Because hair growth makes it difficult, or impossible, to see the areolas, it is important to establish the correct time for the assessment. The presence of nipples/areolae in male pups should be measured when they are obvious (i.e. as they appear in the female litter mates) ideally on PND 12 or 13 (but this may vary with strain); as far as possible, all pups should be evaluated on the same postnatal day as there can be marked differences as maturation progresses. Further guidance on assessment of nipple retention is provided in GD 43 (OECD 2008, paragraph 91).

Data analysis, sensitivity/power

37. When examining nipple retention in a study, the nipples could either be recorded as a yes/no answer or by counting the number of nipples. Nipple retention is a yes/no endpoint if it is expressed as the number of males with or without nipple, but this endpoint can also be semi-quantitative, if the number of nipples is recorded (i.e. from 0 to 12).

38. If only a yes/no answer are used then the power is similar to assessment of malformations. Power calculations illustrate that the effect size needed for detection of quantal effects has to be 25-37% with 20 litters per group and 50-75% with 8 litters per group. This indicates that the sensitivity for detecting effects based on a yes/no answer is quite low irrespective of the number of litters included.

39. This view is also expressed in OECD GD 151 (OECD 2013) where it is stated that: *A quantitative count in male pups is also recommended as a qualitative assessment only (presence/absence) of nipples/areolae may be rather insensitive particularly when control incidence is high (for examples, see Gray et al, 2009 and Christiansen et al, 2010).*

40. Power Simulations of nipple retention based on nipple counts have been made and are described in appendix 1a. The data are from 20 Copenhagen studies (see para 25). The results show that small NR differences can be detected with 8 litters per group if the control baseline rate in male rats is close to zero. If the control baseline in male rats is higher (i.e. 2 nipples) more than 8 litters per group is needed for detection of small NR differences. However, the background level of NR is normally very low (close to zero) for the 'standard' rat strains (e.g. Wistar, Wistar Han, Sprague Dawley).

Human relevance

41. During the last decade, it has become evident that assessment of both AGD (mentioned above) and NR in rodent offspring can be used as markers of impaired androgen action within the critical programming windows of sexual differentiation (Welsh et al. 2008, Welsh et al. 2010). Both endpoints have been shown to be highly predictive of increased risk of adverse reproductive toxicity effects in rats later in life, including increased incidence of hypospadias and cryptorchidism, decreased penile length and seminal vesicle weight

(Bowman et al. 2003, Christiansen et al. 2008, Welsh et al. 2008), and assessment of both AGD and NR has been recognised for regulatory purposes.

42. Nipple retention or number of nipples are not an observed effect in humans, but the relevance of this endpoint is tied to the cause of this effect, which is the ability of chemicals to impair androgen action during development.

43. Nipple retention is mandatory in OECD TG 443 (Extended one-generation reproductive toxicity study (OECD 2012)) where it is stated: *Moreover a statistically significant change in nipple retention should be evaluated similarly to an effect in AGD as both endpoints indicate an adverse effect of exposure and should be considered in setting a NOAEL (ref. GD 151, OECD 2013).* As the NOAEL can be used as the point of departure for setting safe exposure levels for humans this further supports that effects on NR in experimental animals are of human relevance.

Animal welfare

44. Assessment of NR on PND 12 or 13, if included in the test methods (OECD TG 421/422), requires handling of the pups on this day. The assessment of each pup can be done quickly and gently and is therefore not expected to lead to any animal welfare concerns.

Inclusion of NR in TG 421/422

45. There are standardized OECD test methods for assessing NR and the sensitivity analysis shows that relevant data can be obtained with the number of litters per group in the TGs 421/422. Also, NR is an endpoint whose biology and mode of action are relevant to humans and there are no concerns for animal welfare related to the assessment of this endpoint. This all supports that assessment of NR can be included in TGs 421/422. For the OECD TGs 421/422 an extension of the testing period from postnatal day 4 to 12 or 13, i.e. 9-10 day is necessary as nipple retention has to be assessed on postnatal day 12 or 13.

46. A quantitative count in male pups is required as a qualitative assessment only (presence/absence) of nipples/areolae is regarded as being too insensitive.

47. However, the presence of nipples/areolae in male pups have to be measured when they are obvious (i.e. as they appear in the female litter mates) ideally on PND 12 or 13 (but this may vary with strain). Consequently inclusion of this endpoint in OECD TG 421/422 requires that the observation period is extended from postnatal day 3 to postnatal 12 or 13.

Thyroid hormones

Method

48. At the time in 2007/2008, when TG 407 was updated, the TG 407 validation data was judged insufficient to support inclusion of these particular endpoints as mandatory due to

uncertainty about their sensitivity. Therefore in TG 407 the measurements of thyroid hormones (T3,T4 & TSH serum measurements) is optional as it is stated in the beginning of para 37 that:

> *37. Although in the international evaluation of the endocrine related endpoints a clear advantage for the determination of thyroid hormones (T3, T4) and TSH could not be demonstrated, it may be helpful to retain plasma or serum samples to measure T3, T4 and TSH (optional) if there is an indication for an effect on the pituitary-thyroid axis.*

49. The situation was different when the new guideline for the extended one-generation study was developed in 2010-2011 and assessment of thyroid hormones is included here as mandatory. The method is described in paragraph 54 in TG 443, i.e.:

> *54. Systemic effects should also be monitored in F1 animals. Fasted blood samples from a defined site are taken from ten randomly selected cohort 1A males and females per dose group at termination, stored under appropriate conditions and subjected to standard clinical biochemistry, including the assessment of serum levels for thyroid hormones (T4 and TSH), haematology (total and differential leukocyte plus erythrocyte counts) and urinalysis assessments.*

Data analysis, sensitivity/power

50. Intensive power simulations similarly as for AGD and NR (appendix 1a) have also been performed for TH measurements (Appendix 1b). This analysis was performed considering that the blood samples from the sacrificed pups are pooled for males and females from the same litter (i.e. one measure per litter), and compared to data sets of gender-specific non-pooled measurements. Details about statistical testing and the problematic adjustments to p-values for multiplicity, statistical error rates, the power concept and NOAEL determination can be found in appendix 1b. It is assumed that all statistical analysis is based on data obtained under comparable testing conditions. Data were obtained from US EPA, DTU and others (see details in Appendix 1b).

51. To achieve a better data comparability, all statistical analysis were done on the basis of the coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean of the sample. Each individual CV describes the scatter of T4 responses between litters from the same control group. The variation of CVs across different studies and for different age classes is included in Table 1 in appendix 1b. At PND 3-4 pups, the CVs were in the range of 8 up to 77% with median CVs of 8-31% in the data sets. At PND 14-16, the CVs were generally lower, i.e. they ranged from 4-33% with median CVs of 6- 13%. The revisions in the TGs in relation to assessment of thyroid hormones are based upon the above numbers that showed lower CV values in pups at PND 14-16.

52. The statistical analysis showed that detecting a 20% change in a dosed group, compared to controls T4 levels is not likely with 10 litters per group, as this requires at least 17 litters per group. However, there is a high likelihood for detecting a 30% change with 10 litters per group assuming average data variability (appendix 1b).

Human relevance

53. Thyroid hormones (TH) are needed for proper nerve cell differentiation and proliferation, and normal status of these hormones during early development is therefore crucial. In humans even moderate and transient reductions in maternal T4 levels during pregnancy, may adversely affect the child's neurological development. The consequences can be associated with impaired motor- and neurological function in childhood (Pop *et al.,* 1999; Kooistra *et al.*, 2006; Li *et al.*, 2010). Together this indicates that by measuring thyroid hormones in TG 421/422 as an indication of thyroid disruption could indeed be relevant for human risk assessment.

54. The rat is by far the most used *in vivo* model for investigating the toxicity of chemicals suspected to disrupt the hypothalamic-pituitary-thyroid axis (HPT axis) (EFSA, 2011, 2013). However, the relevance of these toxicological rat experiments for humans has been a subject for debate for many years (Döhler et al., 1979, Jahnke & Choksi, 2004; McClain, 1995). It is well-documented that the general construction of HPT axis is the same in rats and humans (Bianco et al., 2002; Zoeller, et al., 2007). However, it is also well-documented that there are quantitative differences between the HPT axis in the two species. It is generally believed that the rat thyroid gland operates at a higher basal activity level than humans'. This is mainly based on the more active histological appearance of the thyroid follicles in rats compared to primates (McClain, 1995), the lack of a high-affinity transport protein (thyroxine-binding globulin) for thyroid hormone in adult rats (Jahnke & Choksi, 2004) and the lower plasma half-life of the two thyroid hormones (THs) thyroxine (T4) and trioidothyronine (T3) in adult rats versus humans (Bianco et al., 2002; Döhler et al., 1979) which necessitates a relatively higher production and secretion rate of TH from the thyroid follicles in rats to keep circulating TH levels constant. Thyroid disrupting chemicals (TDCs) have been shown to decrease circulating levels of THs via various mechanisms and lead to adverse health consequences such as thyroid follicular cell tumours and impaired cognition and/or motor activity (Miller et al. Zoeller, 2009). It seems widely accepted that the formation of thyroid follicular cell tumours in rats due to prolonged elevation of serum thyrotropin (TSH) in response to chemical exposure is not relevant to humans (Capen, 1997; Dellarco, et al. , 2006; Hurley, et al. , 1998). Developmental neural system impairments caused by TDC exposure appear to be independent of TSH and in many instances result from transient changes in circulating TH levels (Crofton, 2008). Relevance analysis suggests that there is a good degree of interspecies concordance in the mode of actions (MOAs) by which these changes in circulating TH occur and the subsequent impairments in the nervous system development, at least qualitatively, making rat data on this endpoint relevant for human health risk assessment (Crofton & Zoeller, 2005, Lewandowski et al. 2004, Zoeller & Crofton 2005).

Animal welfare

55. If blood samples for assessment of thyroid hormones in adults and pups are taken at termination of the study (i.e. PND 13-14) this leads to no concern for animal welfare, as trunk blood can be collected at the time of sacrifice. In adult animals in TG 407, fasted blood samples are to be used and fasting for 20-24 hours in adults only lead to minor concern for animal welfare. However, these blood samples are proposed only to be taken in TG 421/422 if they have not already been taken in a TG 407 study investigating the same compound. The TGs 421/422 studies include relatively similar number of adult animals as TG 407 (5-8 per dose per sex) and therefore, the overall animal welfare considerations will not increase by this assessment and are evaluated as minor. The pups will not be fasted prior to termination and blood sampling because fasting of such young pups would lead to major concern for animal welfare.

Inclusion of thyroid hormones in TG 421/422

56. There are standardized OECD test methods for assessing thyroid hormones. The performed power analysis indicated that CVs of about 20 % and sometimes higher, will be obtained with the number of litters per group in the TGs 421/422. Lower CVs were seen in pups at PND 14-16 compared to PND 3-7. The statistical analysis showed that there is a high likelihood for detecting a 30% change in T4 level with 10 litters per group assuming average data variability. This may at first glance not seem very sensitive. However due to the previously explained species differences in the thyroid hormone system between rats and human, exposure to thyroid disrupting compounds may lead to a greater response on hormone levels in rats than would be expected in humans, and for example a 30 % change in T4 levels is a realistic finding in rats after exposure to a thyroid disrupting chemical. Therefore the assessment of TH in the TG421/422 is sufficiently sensitive and can provide relevant data. Due to the adverse effects seen in humans after developmental hypothyroidism, this endpoint is of high human relevance and there are no concerns for animal welfare related to the assessment of this endpoint as long as blood sampling is done in animals that are being sacrificed anyway. This all supports that assessment of thyroid hormones can be included in TGs 421/422.

Abnormalities of external genital organs

Method

57. In the current TGs 421/422 it is noted that each litter should be examined as soon as possible after delivery to establish the number and sex of pups, stillbirths, live births, runts (pups that are significantly smaller than corresponding control pups), and the presence of gross abnormalities. Thus assessment of abnormalities of genital organs is already to be done. However, no details with regard to how to do this are included.

58. In TG 443 all selected F1 animals are evaluated around sexual maturity and notes are taken for any abnormalities of genital organs, such as persistent vaginal thread, hypospadias or cleft penis.

59. We have in a recent project investigated sexual development in male rat offspring after *in utero* exposure to the endocrine disrupting anti-androgen procymidone. The main purpose of this study was to investigate whether malformations of the male offspring's genitalia could be scored soon after birth and furthermore to access whether it was possible to score the degree of these malformations early after birth. The results (unpublished) presented in Appendix 2 shows that malformations of male offspring's genitalia could be scored early after birth (day 0 and day 6). Also, categorisation of the alterations based on the severity of the effect was possible.

Data analysis, sensitivity/power,

60. We have calculated the effect size needed for finding significant effect, i.e. $p < 0.05$, for yes/no endpoints (Table 2). This was done for studies with 8 or 20 litters per group. As the evaluation may be done in more than one offspring per litter, the calculations also illustrate the effect sizes needed when 2 or 5 offspring per litter is assessed. However, the correct effects sizes needed for 2 or 5 animals per litter are likely to be higher than the ones shown as our calculation is based on the single pup as the statistical unit. To be correct, the calculations should be based on the litter as the statistical unit, i.e. the method should have corrected for litter effects. This was unfortunately not possible for us as there are, to our knowledge, no available easily used statistical programs for that purpose for quantal data.

Table 2. Effect sizes for quantal endpoints needed for p value < 0.05 in one-tailed Fisher Exact test*

Litters per group	Pups per litter	Group	No. with effect	No. without effect	Effect size
20	1	Control	0	20	**0%**
20	1	Exposed	5	15	**25%**
20	2	Control	0	40	**0%**
20	2	Exposed	5	35	**13%**
20	5	Control	0	100	**0%**
20	5	Exposed	5	95	**5%**
8	1	Control	0	8	**0%**
8	1	Exposed	4	4	**50%**
8	2	Control	0	16	**0%**
8	2	Exposed	5	11	**31%**
8	5	Control	0	40	**0%**
8	5	Exposed	5	35	**13%**

*The statistics used when more than one male pup per litter is included is based on using the pup as the statistical unit. Generally, the litter is considered as the correct statistical unit in developmental toxicity studies and using this approach will in most cases lead to even higher effect sizes than those shown in the table.

61. The results in table 2 indicate that for achieving a statistically significant effect with 20 litters per group the frequency of effect in the exposed group has to be 25% with 1 male per litter and 5% with 5 males per litter. With 8 litters per group the frequency of effect in the exposed group has to be 50% with 1 male per litter and 13% with 5 males per litter. These data strongly support that all male pups need to be evaluated, similarly as in OECD TG 414.

62. This limited sensitivity for detecting significant effects on rare adverse outcomes is generally recognized for malformations. Thus, the occurrence of a few similar rare malformations such as hypospadias may generally be considered toxicologically relevant although the finding is not statistically significant.

Human relevance

63. In humans recent studies have reported shorter AGD in boys with hypospadias or cryptorchidism as compared with boys with normal genitalia (Hsieh et al. 2008). Moreover, it is well documented that the incidences of cryptorchidism, hypospadias and testicular cancer have increased over the last decades (Giwercman et al. 1993; Skakkebaek et al. 2001; Boisen et al. 2005).

64. Hypospadias in humans is one of the most common urogenital congenital anomalies affecting boys (Harris 1990). Prevalence estimates in Europe range from 4 to 24 per 10,000 births, depending on definition (Dolk et al. 2004) with higher rates of about 5% reported in a Danish study (Boisen et al. 2005). Little is known about the aetiology of hypospadias, but a

role for EDCs has been proposed, and especially the anti-androgenic EDCs (Baskin et al. 2001).

65. Exposure during critical developmental phases such as *in utero* and in the early postnatal period may lead to adverse effects on both reproductive development and neurodevelopment. The fact that many of the basic mechanisms underlying this developmental process are similar in all mammals indicates that chemicals that have adverse effects on reproductive development in rodents should be considered as potential human reproductive toxicants as well (Gray 1992).

Animal welfare

66. Assessment of abnormalities of external genital organs requires slightly more handling of the new-borns. This assessment can be done very gently and is therefore not expected to lead to any animal welfare concerns. If the assessment is done on PND 12 or 13 prior to termination of the pups, this can similarly be done very quickly and gently and is therefore not expected to lead to any animal welfare concerns. If the assessment of abnormalities of external genital organs is done after termination of the pups on PND 12 or 13, there will obviously be no concern for animal welfare.

Inclusion of abnormalities of external genital organs in TG 421/422

67. Assessment of abnormalities is already included in TG 421/422. However, no details with regard to assessment of abnormalities of external genitals organs are included. The text proposed to be added in the revised TG 421 and 422 in relation to abnormalities is modified from para 30 in OECD TG 414.

Overall discussion and conclusions

68. The aim of this project was to do a feasibility study for minor enhancements of TG 421/422 with ED-relevant endpoints. The endpoints considered for inclusion are anogenital distance (AGD), nipple retention (NR), thyroid hormones and malformations of external reproductive organs in male offspring.

69. For all endpoints, OECD test methods are available for assessing these. Power analyses have been done showing sufficient sensitivity to get relevant data with the number of litters per group in the TGs 421/422. All four endpoints are of relevance for humans as described in this review. All four of them are mandatory to assess in some OECD Test guidelines used for human risk assessment of chemicals. The overall animal welfare considerations will not increase by the assessments of the 4 endpoints. Inclusion of all four endpoints in TG 421/422 does not trigger any animal welfare concerns.

70. In appendix 1a sensitivity of AGD versus Nipple retention is compared in table 4. Most often the sensitivity between NR and AGD is equal (13 studies) and only seldom is AGD more sensitive than NR (2 studies). However, in almost 30% of the studies (6 out of 21)

is nipple retention more sensitive than AGD. Therefore, inclusion of both AGD and nipple retention will provide an increased ability for evaluating the potential endocrine disrupting activity of a substance compared to having only data for AGD. This is especially relevant in cases where equivocal AGD data are found. For the OECD TGs 421/422 an extension of the testing period from postnatal day 4 to 12 or 13, i.e. 9-10 day is necessary as nipple retention has to be assessed on postnatal day 12 or 13.

71. The two Test Guidelines have been updated with specific text proposals. Only minor changes in study design and only few text changes are necessary to include the assessment of anogenital distance (AGD), Nipple Retention (NR), thyroid hormones and malformations of external reproductive organs in TG 421/422.

72. In conclusion, it is feasible to make the proposed minor enhancements of TG 421/422 with ED-relevant endpoints: anogenital distance (AGD), nipple retention (NR), thyroid hormones and malformations of external reproductive organs in male offspring.

References

Baskin LS, Himes K, Colborn T. (2001), Hypospadias and endocrine disruption: is there a connection? Environmental Health Perspectives 109:1175-1183.

Bianco, A. C., Salvatore, D., Gereben, B., Berry, M. J., & Larsen, P. R. (2002), Biochemistry, cellular and molecular biology, and physiological roles of the iodothyronine selenodeiodinases. Endocrine Reviews, 23(1), 38–89.

Boisen KA, Chellakooty M, Schmidt IM, Kai CM, Damgaard IN, Suomi AM, Toppari J, Skakkebaek NE, Main KM. (2005), Hypospadias in a cohort of 1072 Danish newborn boys: prevalence and relationship to placental weight, anthropometrical measurements at birth, and reproductive hormone levels at three months of age. The Journal of Clinical Endocrinology & Metabolism 90:4041-4046.

Bornehag CG1, Carlstedt F, Jönsson BA, Lindh CH, Jensen TK, Bodin A, Jonsson C, Janson S, Swan SH. (2014), Prenatal Phthalate Exposures and Anogenital Distance in Swedish Boys. Environ Health Perspect. 2014 Oct 29.

Bowman CJ, Barlow NJ, Turner KJ, Wallace DG, Foster PMD. 2003. Effects of in utero exposure to finasteride on androgen-dependent reproductive development in the male rat. Toxicological Sciences 74:393-406.

Capen, C. (1997). Mechanistic data and risk assessment of selected toxic end points of the thyroid gland. Toxicologic Pathology, 25(1), 39–48. Retrieved from http://tpx.sagepub.com/content/25/1/39.short

Christiansen S, Scholze M, Axelstad M, Boberg J, Kortenkamp A, Hass U. Combined exposure to anti-androgens causes markedly increased frequencies of hypospadias in the rat. International Journal of Andrology 2008;31:241–8.

Christiansen, S., Boberg, J., Axelstad, M., Dalgaard, M., Vinggaard, A. M., Metzdorff, S. B., and Hass, U. (2010). Low-dose perinatal exposure to di(2-ethylhexyl) phthalate induces anti-androgenic effects in male rats. Reprod. Toxicol 30(2), 313-321.

Clark R, Antonello JM, Grossman SJ, Wise LD, Anderson C, Bagdon WJ, et al. External genitalia abnormalities in male rats exposed in utero to finasteride, a 5 alpha-reductase inhibitor. Teratology 1990;42:91–100.

Crofton, K. M. (2008), Thyroid disrupting chemicals: mechanisms and mixtures. International Journal of Andrology, 31(2), 209–23. doi:10.1111/j.1365-2605.2007.00857.x

Crofton, K. M., & Zoeller, R. T. (2005), Mode of action: neurotoxicity induced by thyroid hormone disruption during development--hearing loss resulting from exposure to PHAHs. Critical Reviews in Toxicology, 35(8-9), 757–69. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16417043

Dellarco, V. L., McGregor, D., Berry, S. C., Cohen, S. M., & Boobis, A. R. (2006). Thiazopyr and thyroid disruption: case study within the context of the 2006 IPCS Human

Relevance Framework for analysis of a cancer mode of action. Critical Reviews in Toxicology, 36(10), 793–801. doi:10.1080/10408440600975242

Dolk H, Vrijheid M, Scott JE, Addor MC, Botting B, de Vigan C, de Walle H, Garne E, Loane M, Pierini A, Garcia-Minaur S, Physick N, Tenconi R, Wiesel A, Calzolari E, Stone D. 2004. Toward the effective surveillance of hypospadias. Environmental Health Perspectives 112:398-402.

Döhler, K. D., Wong, C. C., & von zur Mühlen, A. (1979). The rat as model for the study of drug effects on thyroid function: consideration of methodological problems. Pharmacology & Therapeutics. Part B: General & Systematic Pharmacology, 5(1-3), 305–18. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/386373.

ECHA (European Chemicals Agency). 2013. Evaluation of new scientific evidence concerning dinp and didp in relation to entry 52 of annex xvii to reach regulation (ec) no 1907/2006. Available: http://echa.europa.eu/documents/10162/31b4067e-de40-4044-93e8-9c9ff1960715 [accessed 18 November 2014]

EFSA. (2011). Scientific Opinion on Polybrominated Diphenyl Ethers (PBDEs) in Food. EFSA Journal 2011 (Vol. 9, p. 274 pp.). doi:10.2903/j.efsa.2011.2156

EFSA. (2013). Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile. EFSA Journal 2013 (Vol. 11, p. Appendix F.1 & F.2). Retrieved from http://www.efsa.europa.eu/fr/efsajournal/pub/3293.htm

Eisenberg ML, Hsieh MH, Walters RC, Krasnow R, Lipshultz LI. 2011. The relationship between anogenital distance, fatherhood, and fertility in adult men. PLoS One 6(5):e18973.

Eisenberg ML, Jensen TK, Walters RC, Skakkebaek NE, Lipshultz LI. 2012. The relationship between anogenital distance and reproductive hormone levels in adult men. J Urol 187(2):594-598.

Giwercman A, Carlsen E, Keiding N, Skakkebaek NE. 1993. Evidence for increasing incidence of abnormalities of the human testis: a review. Environmental Health Perspectives 101:65-71.

Gray LE. 1992. Chemical-induced alterations of sexual differentiation: A review of effects in humans and rodents. In: Chemically induced alterations in sexual and functional development: the wildlife/human connection (Colborn T, Clement C, eds). Princeton, NJ:Princeton Scientific Publishing, 203-230.

Gray L, Wolf C, Lambright C, Mann P, Price M, Cooper RL, et al. Administration of potentially antiandrogenic pesticides (procymidone, linuron, iprodione, chlozolinate, p,p_-DDE, and ketoconazole) and toxic substances (dibutyl- and diethylhexyl phthalate, PCB 169, and ethane dimethane sulphonate) during sexual differentiation produces diverse profiles of reproductive malformations in the male rat. Toxicol Ind Health 1999;15:94–118.

Gray, L. E., Jr., Barlow, N. J., Howdeshell, K. L., Ostby, J. S., Furr, J. R., and Gray, C. L. (2009). Transgenerational effects of Di (2-ethylhexyl) phthalate in the male CRL:CD(SD) rat: added value of assessing multiple offspring per litter. Toxicol Sci 110 (2), 411-425.

Harris EL. 1990. Genetic epidemiology of hypospadias. Epidemiologic Reviews 12:29-40.

Hass, U., Scholze, M., Christiansen, S., Dalgaard, M., Vinggaard, A.M., Axelstad, M., et al., 2007. Combined Exposure to Anti-Androgens Exacerbates Disruption of Sexual Differentiation in the Rat, Environ. Health Perspect. 115, 122-128.

Hsieh MH, Breyer BN, Eisenberg ML & Baskin LS 2008 Associations among hypospadias, cryptorchidism, anogenital distance, and endocrine disruption. Current Urology Reports 9 132–142. (doi:10.1007/s11934- 008-0025-0)

Hsieh MH, Eisenberg ML, Hittelman AB, Wilson JM, Tasian GE, Baskin LS. 2012. Caucasian male infants and boys with hypospadias exhibit reduced anogenital distance. Human reproduction 27(6):1577-1580.

Hurley, P. M., Hill, R. N., & Whiting, R. J. (1998). Mode of carcinogenic action of pesticides inducing thyroid follicular cell tumors in rodents. Environmental Health Perspectives, 106(8), 437–45.

Imperato-McGinley J, Binienda Z, Arthur A, Mininberg DT, Vaughan ED Jr, Quimby FW. 1985. The development of a male pseudohermaphroditic rat using an inhibitor of the enzyme 5 alpha-reductase. Endocrinology 116:807-812.

Imperato-McGinley J, Binienda Z, Gedney J, Vaughan ED Jr. 1986. Nipple differentiation in fetal male rats treated with an inhibitor of the enzyme 5 alpha-reductase: definition of a selective role for dihydrotestosterone. Endocrinology 118:132-137.

Jahnke, G., & Choksi, N. (2004). Thyroid toxicants: assessing reproductive health effects. Environmental Health, 112(3), 363–368. doi:10.1289/ehp.6637

Jain VG. Singal AK. 2013. Shorter anogenital distance correlates with undescended testis: a detailed genital anthropometric analysis in human newborns. Human Reproduction 28(9): 2343–2349.

Kooistra L, Crawford S, van Baar AL, Brouwers EP, Pop VJ. 2006. Neonatal effects of maternal hypothyroxinemia during early pregnancy. Pediatrics. 2006 Jan;117(1):161-7.

Kratochwil K. 1971. In vitro analysis of the hormonal basis for the sexual dimorphism in the embryonic development of the mouse mammary gland. J Embryol Exp Morphol. 1971 Feb;25(1):141-53.

Lewandowski, T. A., Seeley, M. R., & Beck, B. D. (2004). Interspecies differences in susceptibility to perturbation of thyroid homeostasis: a case study with perchlorate. Regulatory Toxicology and Pharmacology : RTP, 39(3), 348–62. doi:10.1016/j.yrtph.2004.03.002

Li Y, Shan Z, Teng W, Yu X, Li Y, Fan C, Teng X, Guo R, Wang H, Li J, Chen Y, Wang W, Chawinga M, Zhang L, Yang L, Zhao Y, Hua T. 2010 Abnormalities of maternal thyroid function during pregnancy affect neuropsychological development of their children at 25-30 months. Clin Endocrinol (Oxf). 2010 Jun;72(6):825-9. doi: 10.1111/j.1365-2265.2009.03743.x.

Macleod DJ, Sharpe RM, Welsh M, Fisken M, Scott HM, Hutchison GR, Drake AJ, van den Driesche S. 2010.Androgen action in the masculinization programming window and development of male reproductive organs. Int J Androl. 2010 Apr;33(2):279-87. doi: 10.1111/j.1365-2605.2009.01005.x. Epub 2009 Nov 30

McClain, R. M. (1995). Mechanistic considerations for the relevance of animal data on thyroid neoplasia to human risk assessment. Mutation Research, 333(1-2), 131–42. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8538620

McIntyre BS, Barlow NJ, Wallace DG, Maness SC, Gaido KW, Foster PMD.2000 Effects of in utero exposure to linuron on androgen-dependent reproductive development in the male Crl:CD(SD)BR rat. Toxicol Appl Pharmacol;167:87–99.

McIntyre BS, Barlow NJ, Foster PM.2002 Male rats exposed to linuron in utero exhibit permanent changes in anogenital distance, nipple retention, and epididymal malformations that result in subsequent testicular atrophy.Toxicol Sci. Jan;65(1):62-70.

Mendiola J, Stahlhut RW, Jørgensen N, Liu F, Swan SH. 2011. Shorter anogenital distance predicts poorer semen quality in young men in rochester, new york. Environ Health Perspect 119(7):958.

Miller, M. D., Crofton, K. M., Rice, D. C., & Zoeller, R. T. (2009). Thyroid-disrupting chemicals: interpreting upstream biomarkers of adverse outcomes. Environmental Health Perspectives, 117(7), 1033–41. doi:10.1289/ehp.0800247

Mylchreest E, Sar M, Cattley RC, Foster PM. Disruption of androgen-regulated male reproductive development by di(n-butyl) phthalate during late gestation in rats is different from flutamide. Toxicol Appl Pharmacol 1999;156:81–95.

OECD (2006). Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-Day Oral Toxicity Study in Laboratory Rats. No.59.

OECD (2008). Guidance document on mammalian reproductive toxicity testing and assessment. OECD Series on Testing and Assessment no. 43. Organisation for Economic Cooperation and Development, Paris. 88 pp

OECD (2012). Guidance Document on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption. Series on Testing and Assessment No. 150, ENV/JM/MONO(2012)22

OECD (2013). Guidance document in support of the test guideline on the extended one generation reproductive toxicity study No. 151. Organisation for Economic Cooperation and Development, Paris.

OECD (2015a). Reproduction/Developmental Toxicity Screening Test. Guideline for the Testing of Chemicals (No. 421). Organisation for Economic Cooperation and Development, Paris

OECD (2015b). Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test. Guideline for the Testing of Chemicals (No. 422). Organisation for Economic Cooperation and Development, Paris

Pop VJ, Kuijpens JL, van Baar AL, Verkerk G, van Son MM, de Vijlder JJ, Vulsma T, Wiersinga WM, Drexhage HA, Vader HL. 1999. Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy. Clin Endocrinol (Oxf). 1999 Feb;50(2):149-55.

Salazar-Martinez E, Romano-Riquer P, Yanez-Marquez E, Longnecker M, Hernandez-Avila M. 2004. Anogenital distance in human male and female newborns: a descriptive, cross-sectional study. Environmental Health: A Global Access Science Source 3:8.

Sharpe RM. 2005. Phthalate exposure during pregnancy and lower anogenital index in boys: Wider implications for the general population? Environ Health Perspect 113(8):A504-5

Skakkebaek NE, Rajpert-De ME, Main KM. 2001. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects. Human Reproduction 16:972-978.

Thankamony A, Lek N, Carroll D, Williams M, Dunger DB, Acerini CL et al. 2013. Anogenital distance and penile length in infants with hypospadias or cryptorchidism: Comparison with normative data. Environ Health Perspect http://dx.doi.org/10.1289/ehp.1307178.

van den Driesche S, Scott HM, MacLeod DJ, Fisken M, Walker M, Sharpe RM. 2011 Relative importance of prenatal and postnatal androgen action in determining growth of the penis and anogenital distance in the rat before, during and after puberty. Int J Androl. Dec;34(6 Pt 2):e578-86. doi: 10.1111/j.1365-2605.2011.01175.x. Epub 2011 Jun 2.

Welsh M, Saunders P, Fisken M, Scott HM, Hutchison GR, Smith L & Sharpe RM 2008 Identification in rats of a programming window for reproductive tract masculinization, disruption of which leads to hypospadias and cryptorchidism. J.Clin.Invest. 118 1479-1490.

Welsh M, MacLeod DJ, Walker M, Smith LB & Sharpe RM 1-2-2010 Critical androgen-sensitive periods of rat penis and clitoris development. International Journal of Andrology 33 e144-e152.

Zoeller, R. T., Tan, S. W., & Tyl, R. W. (2007). General background on the hypothalamic-pituitary-thyroid (HPT) axis. Critical Reviews in Toxicology, 37(1-2), 11–53. doi:10.1080/10408440601123446.

Zoeller, R. T., & Crofton, K. M. (2005). Mode of action: developmental thyroid hormone insufficiency--neurological abnormalities resulting from exposure to propylthiouracil. Critical Reviews in Toxicology, 35(8-9), 771–81. **Error! Hyperlink reference not valid.**

Zoeller, R.T. & Crofton, K.M., 2005. Mode of action: developmental thyroid hormone insufficiency--neurological abnormalities resulting from exposure to propylthiouracil. Crit. Rev. Toxicol., 35(8-9), pp.771–81.

Appendix 1a. Power Simulations of Nipple Retention and Anogenital Distance of Rodents exposed to Endocrine Disrupting Chemicals

Objective

The objective of this power simulation study was to determine the allocation of animal numbers per dose in order to study the effect of certain endocrine disrupting chemicals on two endpoints in rodents, nipple retention (NR) and anogenital distance (AGD). The conclusions reached are summarized next. Following that is a detailed justification for these conclusions, including an introduction of key statistical concepts used in this report, the assumptions and constraints on that study, and a description of the data used as basis for the simulations. All analysis is based on data obtained under similar testing conditions, as outlined in the main report.

Conclusions (always for male pups):

1. Litter variability is an important factor in NR and AGD, and data analysis has to account for it.
2. Body weight of the pup is an important co-factor in analysing AGD differences. Its mathematical cubic-root transformation ensures a linear relationship to the measured AGD values.
3. The average AGD size in controls can have an impact on power and the sensitivity of the study design.
4. The sensitivity for detecting AGD differences depends on which minimum AGD difference is considered as toxicological significant, and therefore on how AGD values are normalized to a relative scale.
5. Assuming AGD is scaled to the means from both genders, the detection of a 10% reduction can be ensured only at high litter numbers.
6. Intra-litter correlation for NR can be very low, in extreme cases such that litter can be ignored in data analysis.
7. The closer the control baseline rate for NR is to zero, the higher the statistical power is to identify very small increases in nipple numbers.
8. Small NR differences can be detected at low litter sizes and sufficiently low error rates if the control baseline rate is close to zero.
9. Litter is the statistical unit for designing an experimental study, pup is the statistical unit for data analysis. This does not mean that litter should be neglected in data analysis, but it means that the statistical method should be chosen such that intra-litter variation is reflected in the mean effect estimation.
10. Litter means should not be used in data analysis, but always the pup information.
11. Reducing the litter size to subsamples can reduce the power dramatically.

Description of Data

The majority of data was provided from the same lab (Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark) and was produced over a period of ten years under various different experimental setups in terms of litter numbers, dose numbers, and compounds. Here data were available for both endpoints from 11 independent studies, and as in some studies more than one compound were tested, 22 data sets were considered for data analysis. It should be noted that this comprises not only single chemicals, but also well-defined mixtures, as documented in Table 1. The experimental design from some studies were optimized for regression modelling and thus contained high effect doses which were considered as not relevant for this report and excluded from all data analysis (in these studies always more than three treatment doses were used). If not otherwise stated, data from a minimum of three treatment doses were available, and for NR only data sets were considered if at least one nipple in each treatment group were measured. Due to early data availability all power simulations were based on information from this lab only, and outcomes were then compared and assessed with data information reported from other labs. In total data from 8 other labs were provided, however mainly only for AGD. Here no positive results for NR were reported, and therefore considered as not relevant for data analysis and informative for power simulations. Often only litter means were reported, with no details on how the means were calculated or how many pups were measured, and therefore these data sets were not considered for power analysis. The lack of external data for NR must therefore be considered as a relevant constrain.

Data analysis and description followed always the same purpose: if possible, establishing a NOAEL and LOAEL, and providing information relevant for the simulation studies based on the statistical dose-response model and test. The latter involved information about the number of litters, the average litter size, model-relevant information such as estimations about the within- and between-litter variation, the mean estimates for the NOAEL and LOAEL, and post-hoc power analysis.

Endpoint Modelling

Common to both endpoints is that pup information from the same litter is likely to be more similar than from other litters, which has to be accounted in data analysis. Furthermore, AGD is correlated with the body weight of the pup, which also has to be reflected in data analysis. Therefore both endpoints are from a statistical point of view more complex than most commonly used endpoints in toxicology, and no unique approach exists on how to model and analyse them. As consequence not only different methods are available, but the degree of model complexity is also subjective. We chose statistical representations which are well-accepted in the statistical community and most robustness in terms of model assumptions and commercial software availability. For the correlated data structure of NR we favoured the generalized estimating equation (GEE) model which belongs to the class of marginal models (Liang and Zeger, 1986; McCullagh P and Nelder JA, 1989*)*, and for AGD mixed effect

models with litter treated as random effects (Littell et al, 2006; Verbeke G and Molenberghs G, 2000). For each model a higher model complexity was possible and would have resulted occasionally into a better data presentation, however, we consider the data amount available in most studies as not sufficient to justify more model parameters, and to our experience its impact on the NOAEL determination is minimal.

Statistical Testing

In the same way as the statistical modelling of the endpoint can be done in various ways, there is no universal or common approach on how to perform the statistical testing for these endpoints in order to determine a NOAEL or LOAEL. Crucial is that is depends on the endpoint modelling and corresponding model assumptions. The general goal of an appropriate test is usually to combine good power behaviour with an easy numerical implementation of the test statistics (a problem of particular importance for the experimenters) and robustness against specific violations of the test assumptions (e.g., normality). Especially in the last decade powerful approaches have been developed, such as the so-called Multiple contrast tests (Bretz F and Hothorn LA, 2003). These lead to flexible tests that are easy to implement in complex statistical dose-response models and testing scenarios, such as AGD and NR (references our papers). Depending on the expected shape of the dose-response data, contrast coefficients can be chosen such that they follow certain pattern (trend, non-monotony). In this report we used always contrast tests embedded in the chosen statistical model, with pairwise single-contrasts in analogy to the Dunnett test. They make no assumptions about the shape of the dose-response relationship. For more details see Bretz F & Hothorn LA, 2003.

Adjustments to P-values for Multiple Tests

A common problem with comparing more than one treatment group against the same control group is that several statistical tests are done and each having the chance of declaring a difference between treatment and control to be significant when in fact there is no real treatment effect (false positive). Typically, this type of error is set to an acceptance level of α=5%, i.e. if the statistical test responds with a p-level below the pre-defined α it is concluded that the difference observed between control and treatment means is not due to a chance finding. If several tests are done, each with a 5% chance of incorrectly declaring an effect to be significant when no true difference exists, then the chance that at least one of these tests falsely declaring a significant effect has to be higher than 5%. As a consequence, some adjustment is usually made to control the overall chance of at least one of the many tests being wrong.
Some standard statistical tests have built-in adjustments (e.g., Dunnett, Williams and Jonckheere test), however, they cannot apply to more complex endpoints such as correlated endpoints. The simplest approach to maintain an overall false positive rate is to adjust the p-value after the pairwise comparison tests (multiplicity adjustment), and there are several adjustment schemes possible (e.g., Bonferroni, Bonferroni-Holm, Hochberg or Sidak). They

differ in how well they preserve the overall family-wise error (FWE) rate, and can have a huge impact on deciding whether the testing hypothesis of "no treatment" effect can be rejected in favour of a likely treatment effect, or not. Moreover, if additional assumptions such as monotonicity in the dose-response pattern can be made ("trend"), then even more powerful adjustment can be performed (step-down trend procedures). As consequence, the chance of overlooking existing treatment effects is increased, and approaches have been developed which balances better the false-positive and false-negative rates (so-called false discovery rate, FDR).

The following table provides an example about how adjustment procedures can change raw p values:

	Unadjusted p value	Bonferroni adjusted p value	Hochberg adjusted p value	FDR
Control - Treatment 1	0.0130	0.0390	0.0390	0.0390
Control - Treatment 2	0.0325	0.0975	0.0550	0.0488
Control - Treatment 3	0.0550	0.1650	0.0550	0.0550

Often the choice of the adjustment is made on practical constraints, such as software availability, or the data analyst is not aware about this (even often for statisticians) confusing field. As a monotonic trend cannot be guaranteed *a priori* for the endpoints selected in this report, all power analysis was based on unadjusted p values. Depending on how many treatment groups are planned for the study, possible p-value adjustments should be taken in consideration at the planning stage.

Error Rates in Statistical Testing and the Power Concept

Statistical hypothesis tests use data from a sample in order to make inferences about a statistical population. Typically for toxicology, we assume as Null hypothesis "no treatment effect", and the aim of the experimental study is to provide sufficient data evidence for rejecting the Null hypothesis, and as consequence accepting the alternative hypothesis ("treatment effect"). The decision can be done wrongly in two different ways, illustrated in the following table:

		state of the world	
		H_0 is true „no effect“	H_0 is false „dose related effect”
results of hypothesis testing	Accept H_0 (no significance)	1-α (no error)	Type II error β
	Reject H_0 (significance)	Type I error α (significance level)	1-β = power (no error)

The probability of rejecting a true null hypothesis (an effect is accepted as significant, while in truth no effect exists) is the Type I error, also called the false positive error rate. The probability of a Type II error occurring is referred to as the false negative rate (β). Power is equal to 1 − β, which is also known as the sensitivity. Most researchers assess the power of their statistical tests using 0.80 as default, meaning that the probability for a false negative is less than 0.2. This convention implies a four-to-one trade-off between the probability of a Type II error and a Type I error, when α=5% is selected as criterion for statistical significance. Therefore the power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false (i.e. the probability of not committing a Type II error, hence the probability of not making a false negative decision on whether to reject a null hypothesis). In other words, power is the probability of finding a difference that does exist.

Power is a function of α, sample size, the effect difference between control and treatment mean, and the data variation of the endpoint. It is also conditional of the chosen statistical test. Power is strongly influenced by sample size, i.e. if sample sizes are small, the power of any test is usually low, and reducing α reduces always the power, i.e. over-controlling type I error rates increases the chance of false-negative rates. The greater the data variability, the less the statistical power, and the stronger the effect differences of interest, the more likely to detect it. Powerful statistical tests can detect small differences, weak tests only large differences, and the only way to reduce both error rates at the same time is to increase the sample size.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. Although these error rates correspond to long-run outcomes, and therefore no guarantee is given that the actual study will follow exactly the assumptions made in the power analysis, nevertheless one could get a sense of whether the experimental design was a credible one, and whether it is likely to minimize the two kinds of errors that are possible in dose-response data and, correspondingly, maximize the likelihood of making a correct decision.

In this report the two error rates were set to α=5% (two-sided) and β=20%, i.e. we assumed a power of 80% as minimum.

Simulation studies

For each endpoint, the power and sample size estimation should be based on the proposed dose-response model for the endpoint and data of primary interest. Because of the complex statistical nature of both endpoints no exact or approximate mathematical expression exists which determines the exact sample size at given power (or vice versa). Therefore it was necessarily to perform computer-intensive simulation studies, based on the information obtained from all available dose-response data for these endpoints. The power analysis was conducted using Monte Carlo simulation, by simulating a complete dose-response data set in

which the treatment effect is given, and generating numerous samples that have comparable size and variance structure as the actual data. Main assumption for reliable simulation outcomes is that the underlying model is a sufficiently accurate representation of the data and the study design. The probability of getting a statistically significant result from the model is equal to the probability of getting a statistically significant result in the study, that is, the statistical power. The probability of a significant result in the model can be calculated by repeating the simulation a large number of times and computing the proportion of runs that produced significant results. Thus, in order to estimate the statistical power of the test at a given effect difference and experimental setup, we repeated the simulation 5000 times at that effect size, and recorded the proportion of runs that were statistically significant using the test and a significance criterion of α=0.05 (two-sided).

Power simulations can be broken down into three steps: first it requires describing and modelling the underlying distribution from which the data are thought to arise. Most often this involves making assumptions about the distribution based on empirical results from studies that have already been conducted and which share characteristics with the study being planned. Using those data it was able to obtain estimates for the nuisance mean model parameters, variance-covariance matrix of the random effects, and error variance. As variability in the data varies from study to study, we defined an average data scenario mirroring average data variability, and a worst-case data scenario assuming an unlikely (but not unrealistic) high data variability. The latter allows assessing the impact of high data variations on power. The second step is to generate a large number of samples from the assumed true noise distribution using various sample sizes that are thought to be adequate to achieve the desired power, and the third step is to fit the assumed model to the samples that have been generated. For each simulated data set, we perform a hypothesis test and determine if sufficient evidence exists to reject the null hypothesis for that sample. Once all sample data sets have been processed, we can use the testing results to estimate the power of the test. For preliminary simulations where the approximate sample size is not well known, we considered a wide range of sample sizes and use smoothing splines to get an approximation of the power function.

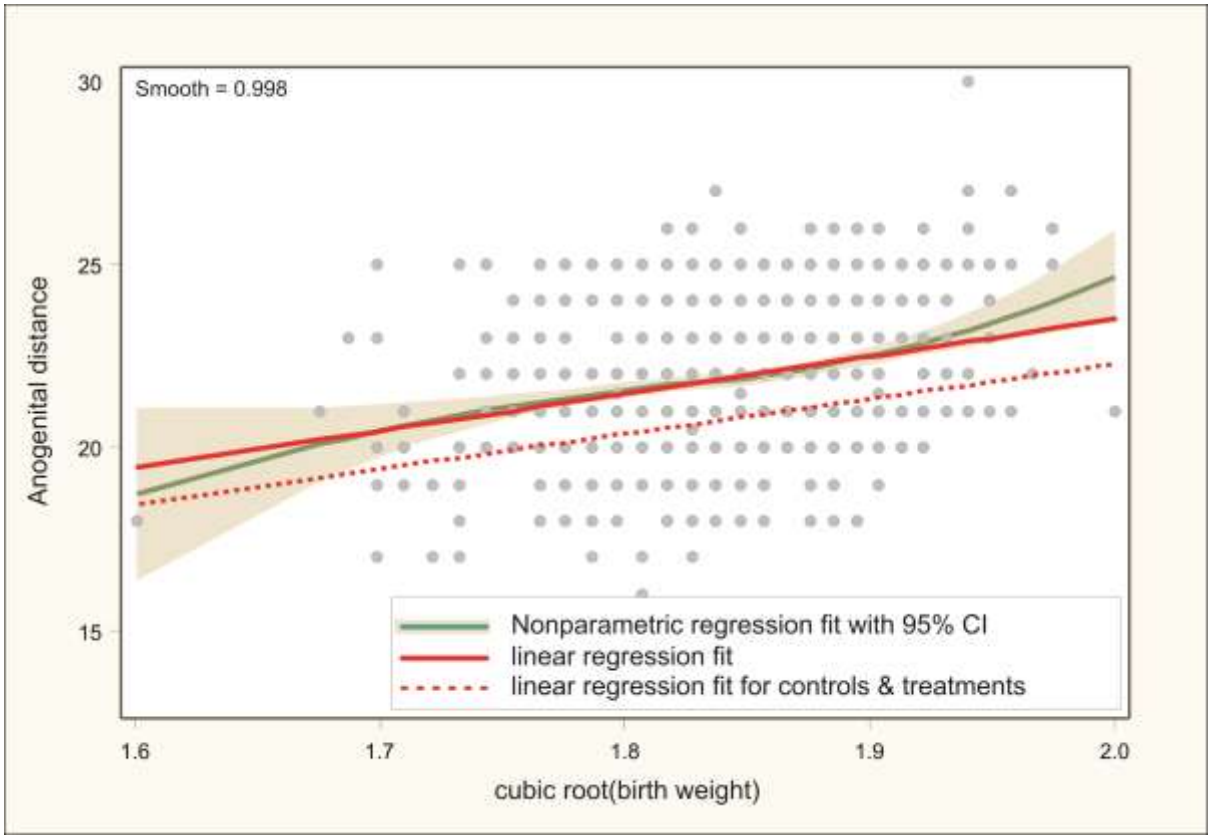
Anogenital Distance (AGD)

Description of Statistical Model and Estimation Method

AGD was analysed by mixed effect modelling (LMM), with litter treated as random effects. This approach simplifies and unifies many common statistical analyses, including those involving repeated measures, random effects, and random coefficients. The basic assumption is that the data are linearly related to unobserved multivariate normal random variables. For that purpose it was necessary to transform body weight such that it could be used as linear co-variable. This was realized by the cube root transformation. The cube root is commonly used because it is thought that this conversion provides the best comparison between the three-dimensional end point (weight) and the one-dimensional AGD. No indications were found against the normality assumption. In theory, control and treatment groups can have different

linear relationship between body weight and their AGD responses. This is certainly justified for gender differences (females have a lower AGD to birth weight ratio), and thus likely to be the similar case for males at high effect doses. However, for moderate treatment responses we found no clear evidence for dose-specific linear relationships between birth weight and AGD, and assumed a treatment-independent relationship which was estimated from each data set. Figure 1 shows from all male controls their individual birth weights and AGD, together with a nonparametric (solid green line) and linear regression fit (solid red line): the agreement between both curves indicates that the linearity assumption is justified, and when the linear regression is repeated including all control and treatment male pups, the corresponding linear regression curve is shifted downwards (dotted red line) without changing significantly its steepness (supporting the assumption of a linear relationship independent of the treatment). The estimated steepness parameters are reported for all data sets in table 1 (tet_{BW}).

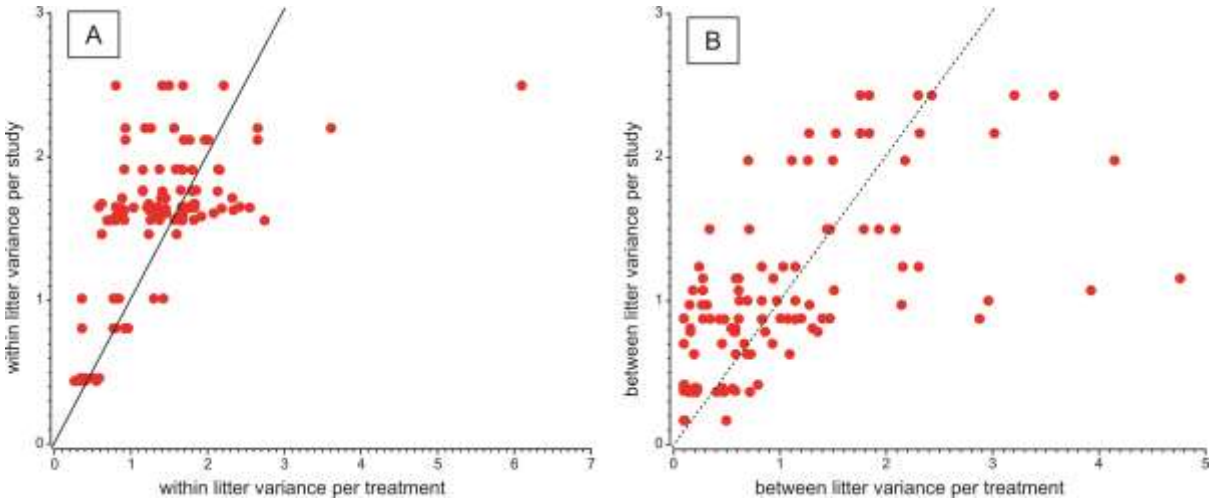
Figure 1: Relationship between birth weight (g) and AGD (units) in male controls



Important for performing simulation studies is knowledge about the inter- and intra-litter variation. This required a model decision about treatment-specific vs. general estimates for these two sources of variation. The first would mean a complex and more accurate modelling step, the last would favour a more robust approach. Therefore we estimated for each data set both model specifications in order to get an impression about the variation of the treatment-specific estimations, here the variances. Figure 2 shows for the within-litter variability (A) and between-litter variability (B) variance estimates for all data sets, with the treatment-specific estimates on the x-axis and the overall study estimate on the y- axis. In both cases the variation around an overall mean estimate was moderate (horizontal data scatter), justifying

the assumption of an overall, treatment-independent estimate for the within-litter and between-litter AGD variability.

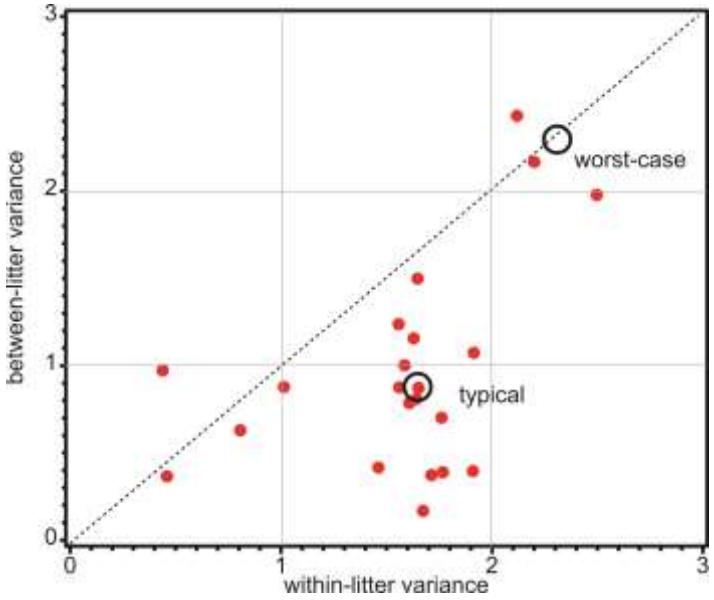
Figure 2: AGD in male pups – treatment-independent study variability vs. treatment variability, shown for within-litter variability (A) and between-litter variability (B).



To define an average and worst-case scenario for the power simulations, all estimates for within-litter and between-litter variances are plotted in Figure 3. The sum of both estimates defines the total variance of the mean AGD estimates. Typically, most dots are below the trend line, suggesting that the main variation of AGD measurements arises from the litter, and not from between litters. For the power simulation studies we set $Variance_{between\ litter}$=0.87 and $Variance_{within\ litter}$ =1.64 for an average data variability scenario, and $Variance_{between\ litter}$=2.3 and $Variance_{within\ litter}$ =2.3 defining a worst-case data variability scenario.

All data analyses were performed by using the MIXED procedure in the statistical software SAS.

Figure 3: Anogenital distance in male pups – within-litter vs. between-litter variability, with circles defining the typical and worst-cast variability settings used for power simulation.



Description of Simulations

Using the model described above, we simulated data for litter sizes from 5 to 20. For each sample size, 5000 samples were generated and analysed using different mean AGD responses for the controls. As effect differences of interest 10% and 20% reduction on the relative AGD scale normalized to both gender controls were selected. The underlying concept behind the simulations is described in detail by Stroup (1999) which uses the Non-Central parameter of a Non-Central F-Distribution to generate correlated random samples according to the pre-defined between-litter and within-litter variances. This reduces significantly the time needed for a single simulation step, and allows the analyses of various experimental setups in a relatively short time.

We used three different scenarios for the litter sizes: either they were hold fixed at 2 and 8, respectively, simulating extreme litter sizes, or in a third scenario they were resampled by random out of a pool from all observed control litter sizes. The resulting variation of this variable litter size setting is summarized by box whisker plots, with boxes representing the quartiles of the simulation outcomes and the whiskers the 5th percentile and the 95th percentiles. Birth weights were resampled with replacement out of a pool of all measured male control pups. Their linear relationships to AGD were defined by setting the corresponding model parameter tet_{BW} to 9.57.

All data analyses were performed by using the IML procedure in the statistical software SAS.

Table 1: AGD – Dose-Response summary (Copenhagen studies)

Study	type of treatment	dose	litter		birth weight		Model information (LMM)				AGD (units)		Control difference			
			#	av. size	mean	90% percentile	between-litter variance	within-litter variance	ILC	tet_{BW}	mean	SEM	abs	ratio	norm	post-hoc power
A	male control		7	5.4	6.36 [5.80-6.80]						21.33	0.698	0	1	1	-
	compound A	NOAEL[1)]	10	4.1	5.98 [5.20-6.60]		1.98	2.50	0.44	9.01**	19.67	0.441	-1.66	0.92	0.84	31.9%
	female control		8	5.2	5.93 [5.40-6.30]						11.08	0.333	-10.25	0.52	0	-
B	male control		13	4.8	6.32 [5.80-7.00]						21.40	0.142	0	1	1	-
	compound B	NOAEL	6	5.8	6.27 [5.80-6.70]		0.63	0.81	0.44	8.91**	21.03	0.389	-0.37	0.98	0.96	12.6%
		LOAEL	6	4.2	6.40 [5.70-6.90]						19.53**	0.371	-1.87	0.91	0.82	98.9%
	compound C	LOAEL[2)]	7	4.0	6.24 [5.70-6.90]		0.88	1.01	0.46	6.24**	19.38**	0.422	-2.02	0.91	0.80	97.9%
	female control		13	5.8	6.01 [5.40-6.60]						11.09	0.112	-10.31	0.52	0	-
C	male control		15	4.7	6.42 [5.35-7.00]						22.48	0.409	0	1	1	-
	compound D	NOAEL	8	4.4	6.41 [5.70-7.00]		2.43	2.12	0.53	14.93**	20.56	0.583	-1.93	0.91	0.82	59.7%
		LOAEL	8	4.9	6.21 [5.80-6.60]						19.17**	0.583	-3.32	0.85	0.68	94.2%
	compound E	LOAEL[2)]	8	4.1	6.55 [6.00-7.10]		2.17	2.20	0.50	12.92**	20.75*	0.537	-1.73	0.92	0.84	72.2%
	female control		13	5.7	6.10 [5.60-6.50]						11.98	0.397	-10.51	0.53	0	-
D	male control		15	4.3	6.23 [5.50-6.90]						20.68	0.325	0	1	1	-
	compound F	NOAEL[1)]	7	4.7	6.20 [5.40-6.70]		1.24	1.56	0.44	9.40**	20.69	0.636	0.01	1	1	31.7%
	compound G	NOAEL[1)]	8	4.8	5.84 [4.80-6.70]		1.00	1.59	0.39	5.32**	19.70	0.408	-0.98	0.95	0.91	24.7%
	compound H	NOAEL	8	6.4	6.23 [5.70-6.90]		0.87	1.65	0.35	7.52**	19.78	0.364	-0.90	0.96	0.91	36.5%
		LOAEL	5	5.2	6.18 [5.50-6.60]						17.96**	0.410	-2.72	0.87	0.74	97.9%
	female control		15	5.1	5.95 [5.30-6.60]						10.32	0.262	-10.36	0.50	0	-
E	male control		13	4.8	6.19 [5.70-6.80]						19.97	0.440	0	1	1	-
	Mixture A	NOAEL	14	5.3	5.85 [5.40-6.40]		1.50	1.65	0.48	5.60**	19.08	0.400	-0.89	0.96	0.90	29.4%
		LOAEL	15	4.7	6.20 [5.40-6.90]						18.23**	0.357	-1.75	0.91	0.81	82.8%
	female control		14	5.5	5.97 [5.50-6.70]						10.68	0.264			0	-
F	male control		13	4.5	6.07 [5.40-6.60]						20.67	0.270	0	1	1	-
	compound I	NOAEL[1)]	7	3.9	6.28 [5.60-6.90]		1.07	1.91	0.36	14.95**	21.25	0.805	0.59	1.03	1.06	12.6%
	compound J	NOAEL[1)]	7	5.0	5.84 [5.30-6.40]		1.16	1.63	0.42	9.06**	20.63	0.445	-0.04	1.00	1.00	20.6%
	Mixture B	LOAEL[2)]	16	5.6	6.19 [5.40-6.80]		0.88	1.56	0.36	4.79*	19.22**	0.192	-1.45	0.93	0.85	95.5%
	female control		13	5.2	5.71 [5.00-6.40]						10.80	0.193			0	-

Study	type of treatment	dose	litter		birth weight		Model information (LMM)				AGD (units)		Control difference			
			#	av. size	mean	90% percentile	between-litter variance	within-litter variance	ILC	tet_{BW}	mean	SEM	abs	ratio	norm	post-hoc power
G	male control		13	5.2	6.25 [5.40-7.00]						20.61	0.139	0	1	1	-
	Mixture C [3]	NOAEL[1]	14	5.9	6.38 [5.80-6.90]		0.17	1.67	0.09	7.57**	20.48	0.245	-0.13	0.99	0.99	16.7%
	female control		14	6.1	6.00 [5.30-6.50]						10.31	0.089			0	-
H	male control		19	5.5	6.29 [5.60-6.90]						21.57	0.215	0	1	1	-
	Mixture D	NOAEL	19	5.2	6.25 [5.70-6.80]		0.79	1.61	0.33	7.71**	21.24	0.253	-0.33	0.98	0.97	12.5%
		LOAEL	14	6.0	6.53 [6.00-7.20]						20.83*	0.242	-0.74	0.97	0.93	59.5%
	Mixture E [3]	LOAEL[2]	15	4.9	6.44 [5.80-7.00]		0.37	1.71	0.18	8.74**	21.06*	0.229	-0.51	0.98	0.95	52.7%
	Mixture F [3]	NOAEL[1]	17	5.2	6.24 [5.60-6.80]		0.81	1.64	0.33	8.46**	21.92	0.321	0.35	1.02	1.03	24.9%
	female control		20	5.9	5.96 [5.10-6.50]						10.87	0.129			0	-
I	male control		18	6.4	6.22 [5.60-7.10]						24.61	0.088	0	1	1	-
	compound K	LOAEL[2]	21	5.7	6.28 [5.50-6.80]		0.37	0.46	0.44	3.64**	24.17*	0.121	-0.44	0.98	0.96	59.4%
	female control		19	5.6	5.85 [5.20-6.50]						13.57	0.110			0	-
J	male control		15	4.9	6.37 [5.70-6.90]						24.00	0.172	0	1	1	-
	compound L	NOAEL	13	6.1	6.43 [5.70-7.00]		0.97	0.44	0.69	8.99**	24.12	0.161	0.13	1.01	1.01	0.8%
		LOAEL	15	6.0	6.59 [6.00-7.10]						22.85**	0.300	-1.14	0.95	0.89	92.8%
	female control		15	5.7	6.11 [5.60-6.60]						13.42	0.116			0	-
K	male control		15	5.0	6.47 [5.70-6.90]						21.98	0.239	0	1	1	-
	Mixture G [3]	LOAEL[2]	16	5.6	6.49 [5.35-7.20]		0.39	1.77	0.18	9.35**	20.50**	0.187	-1.48	0.93	0.86	99.5%
	Mixture H [3]	NOAEL[1]	16	5.0	6.51 [5.60-7.15]		0.70	1.76	0.28	9.66**	21.50	0.282	-0.48	0.98	0.95	19.6%
	Mixture I [3]	LOAEL[2]	17	5.9	6.17 [5.50-6.90]		0.40	1.91	0.17	6.34**	20.73**	0.190	-1.25	0.94	0.88	89.7%
	female control		14	5.4	6.16 [5.50-6.80]						11.37	0.124			0	-

ILC=inter-litter correlation; * stat. significant at α=5%; ** stat. significant at α=1%;

[1)] all doses produced significant responses, values for the lowest dose are shown; [2)] all doses produced non-significant responses, values for the highest dose are shown; [3)] only two treatment doses were tested; tet_{BW} model parameter estimated for cubic root-transformed body weight;

Outcomes of the power simulations

The outcomes of the power simulations are summarized for control pups with high AGD estimates in Figure 4 (males: 24.61 units, females: 13.57 units), and for control pups with a low AGD baseline in Figure 5 (males: 20.00 units, females: 10.70 units). For these two scenarios two treatment-related relative reductions to the male controls were investigated, 10% (A) and 20% reduction (B), with both reductions related to the difference between male and female controls. In Figure 4A the 10% reduction corresponds to a 4.5% reduction in relation to the control male AGD (i.e. ratio between 1.104 and 24.61), while the 10% reduction in Figure 5B corresponds to a 4.65 % reduction in relation to the control male AGD. Similarly for the 20% reductions, they correspond to a 9% reduction (Figure 4B) and 9.3% reduction (Figure 5B), respectively.

If the detection of a 20% reduction at high likelihood is of interest (power > 80%), only small sample sizes up to 9 litters are required, without increasing the false-positive error rate. This assumes that all pup information is used for data analysis. However, smaller effect sizes are likely to be harder to detect, the power analysis suggests that a 10% reduction will only be detected with a sufficient certainty if the AGD information from at least 16 litters is available and the individual variation follows the average pattern. Ideally the control AGD should then be also not too low.

Whereat in the previous figures the effect size of interest was given (10 and 20% reduction, respectively) and the power was estimated in dependence of the sample size (litter numbers), we also analysed the reverse situation by fixing the power and estimating the effect size in dependence of the sample size (litter numbers), i.e. the sensitivity. The results are shown in Figure 6, again for relatively large AGD units in the controls (top figure) and small ones (bottom). Here we focused only on average litter sizes, resampled from all available data sets, but again for a typical data variability scenario (green curve) and a worst-case (red curve). The horizontal lines correspond to the control AGD means. Any effect difference between the male control line and the curves is unlikely to be detected as statistically significant, at least at given 80% power.

Figure 4: Power simulation for male pups with high AGD control estimates. 10% reduction corresponds to a 4.5% reduction in relation to the control male AGD, while the 20% reduction correspond to a 9% reduction in relation to the control male AGD.

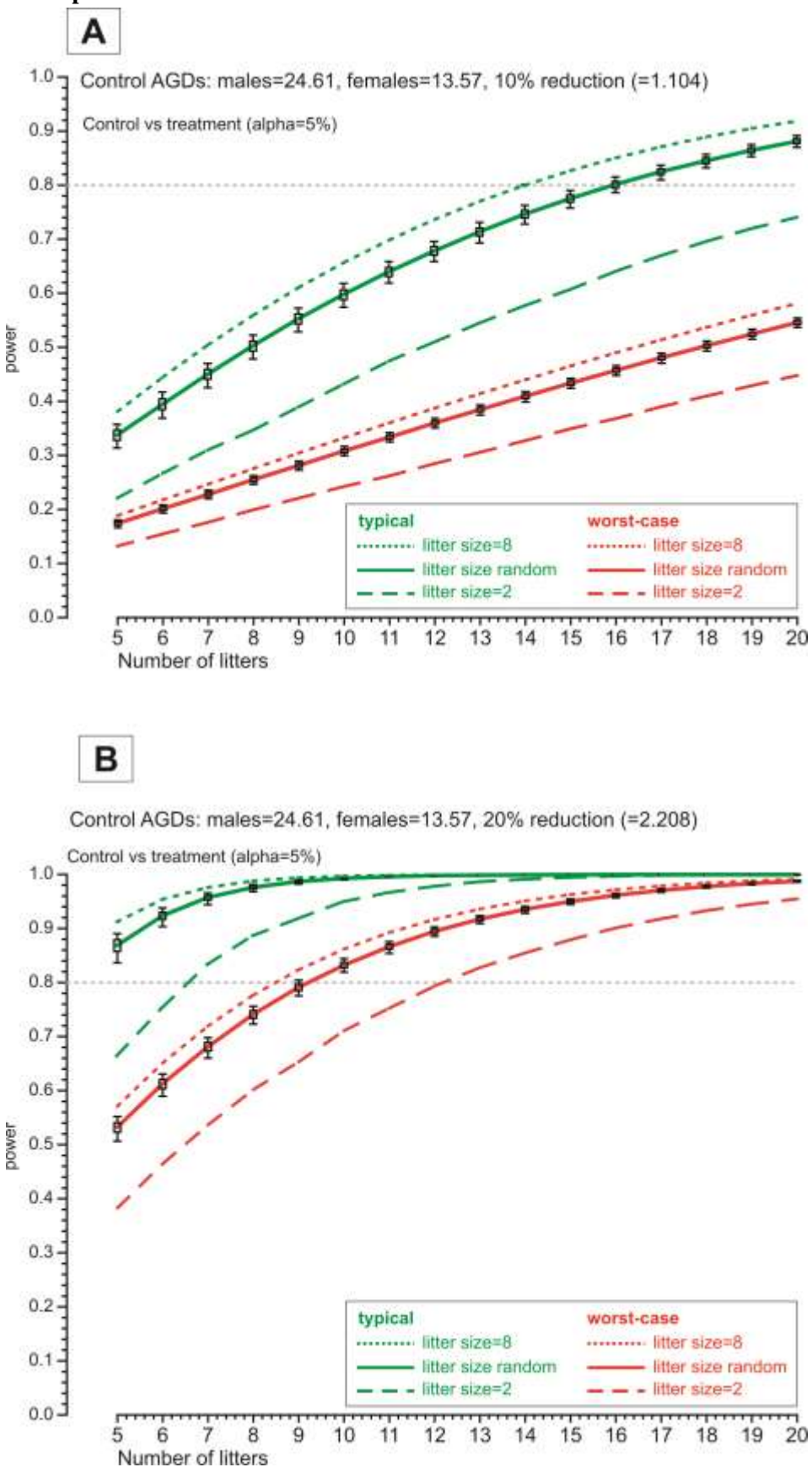


Figure 5: Power simulation for male pups with low AGD control estimates. 10% reduction corresponds to a 4.65% reduction in relation to the control male AGD. 20% reduction corresponds to a 9.3% reduction.

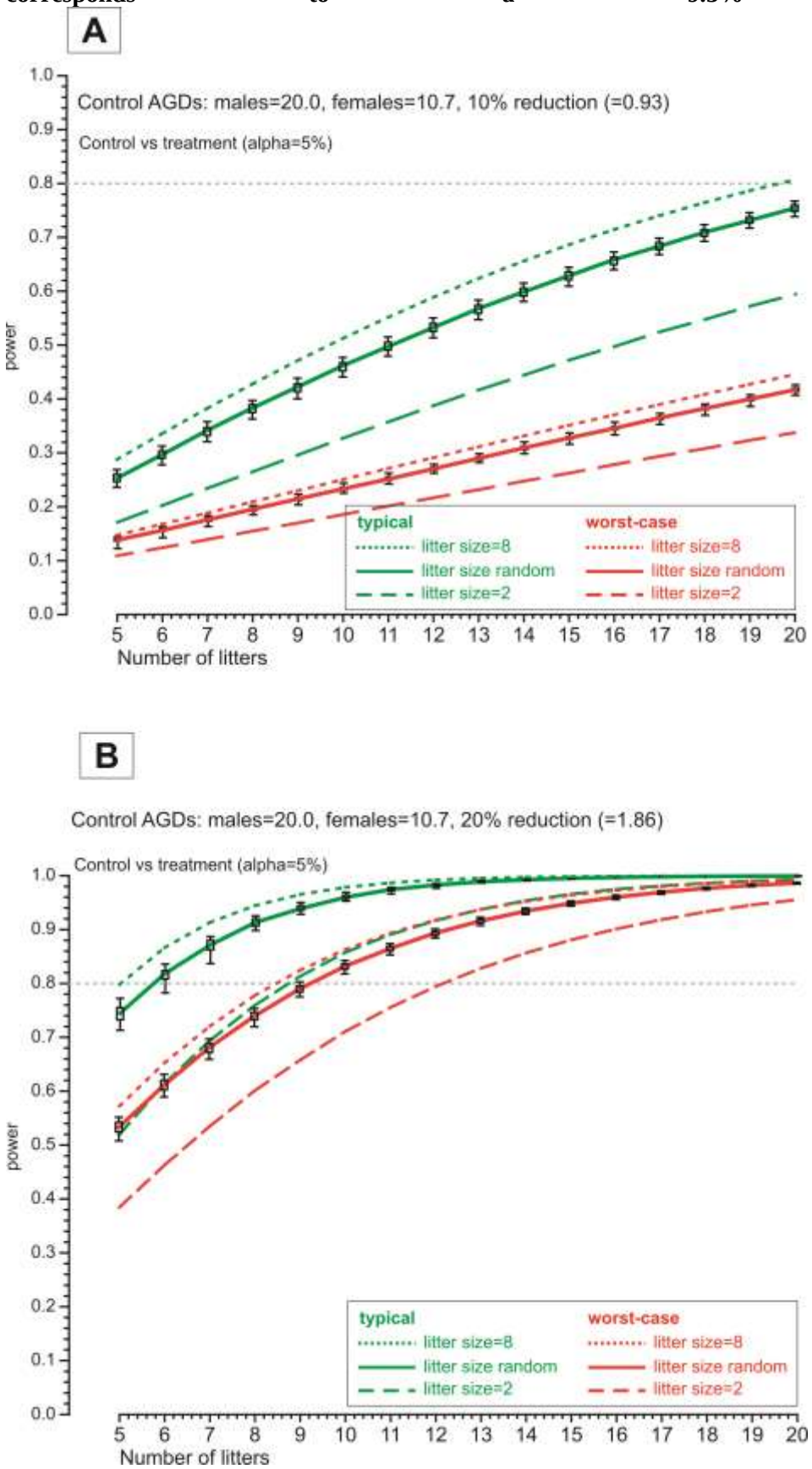
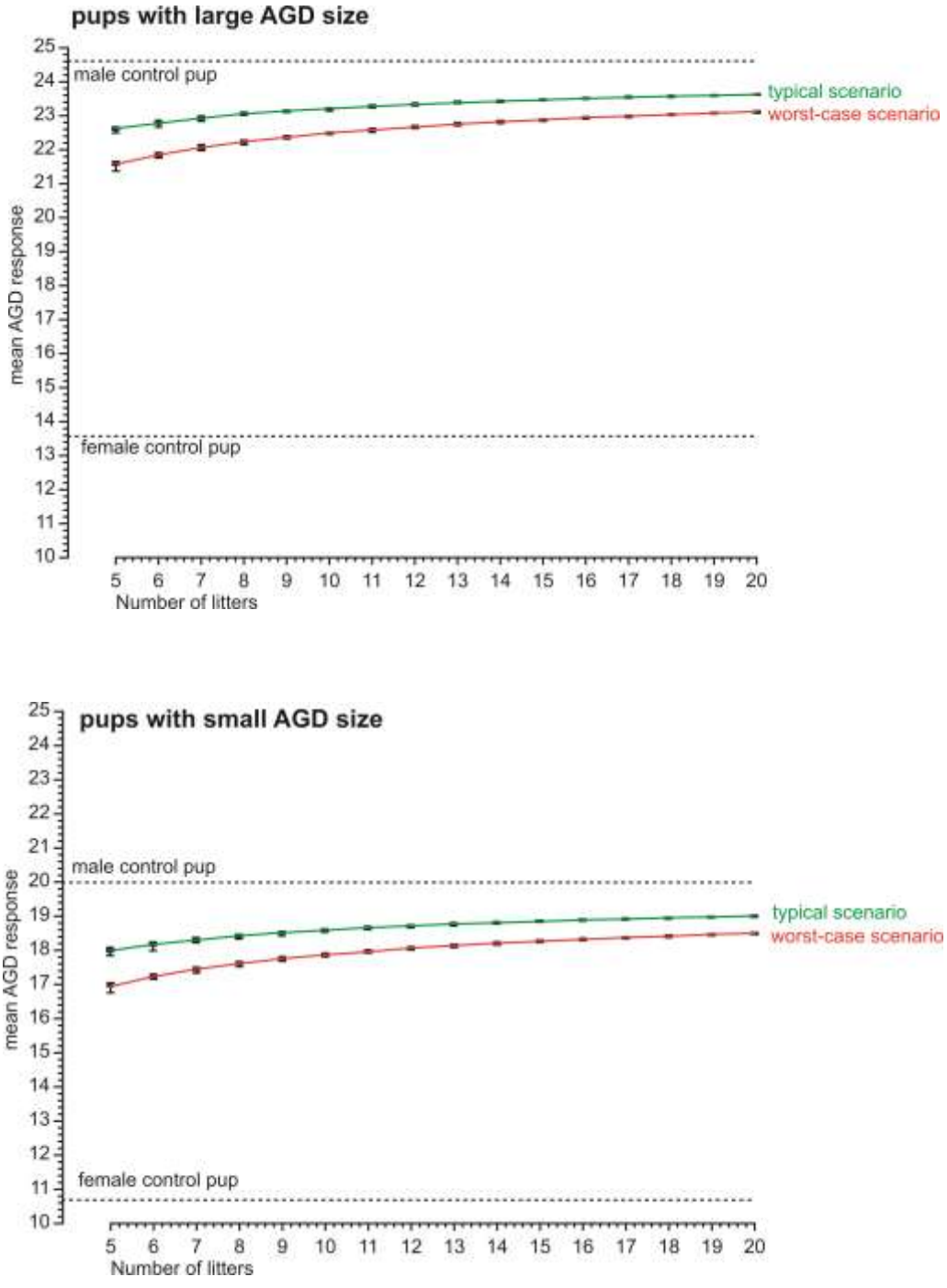


Figure 6: Statistical detection limit (sensitivity) at given power (80%) and false-positive error rate α=5%. Large and small pups refer to the average AGD sizes from the male controls as mentioned in the previous figures.



AGD Data from other labs

Data sets from 7 studies were analysed and outcomes were compared with that from the Copenhagen studies. Only in two studies significant treatment effects were detected, with different directions compared to their controls: study C revealed an increase in AGD, study D a reduction in AGD (Table 2). Most studies were performed with huge numbers of damns, litter numbers ranged from 24 up to 30. The litter sizes and birth weights were comparable to the previous results (exception: study F), however, significant differences were between the mean AGD estimates and their variance components: the between-litter as well as within-litter variances were smaller by a factor of 10-100. The reasons for these gross differences are unknown, but probably indicate that AGD values were reported in a different unit. It should be also noted that AGD values were often reported as rounded values. Based on these values the power simulation studies were repeated, with setting $Variance_{between\ litter}$=0.04 and $Variance_{within\ litter}$ =0.06 for an average data variability scenario, and $Variance_{between\ litter}$=1.5 and $Variance_{within\ litter}$ =1.5 for a worst-case data variability scenario, results are shown in Figure 7. Here a 10% reduction corresponds to a 5% reduction in relation to the control male AGD, while the 20% reduction corresponds to a 10% reduction in relation to the control male AGD.

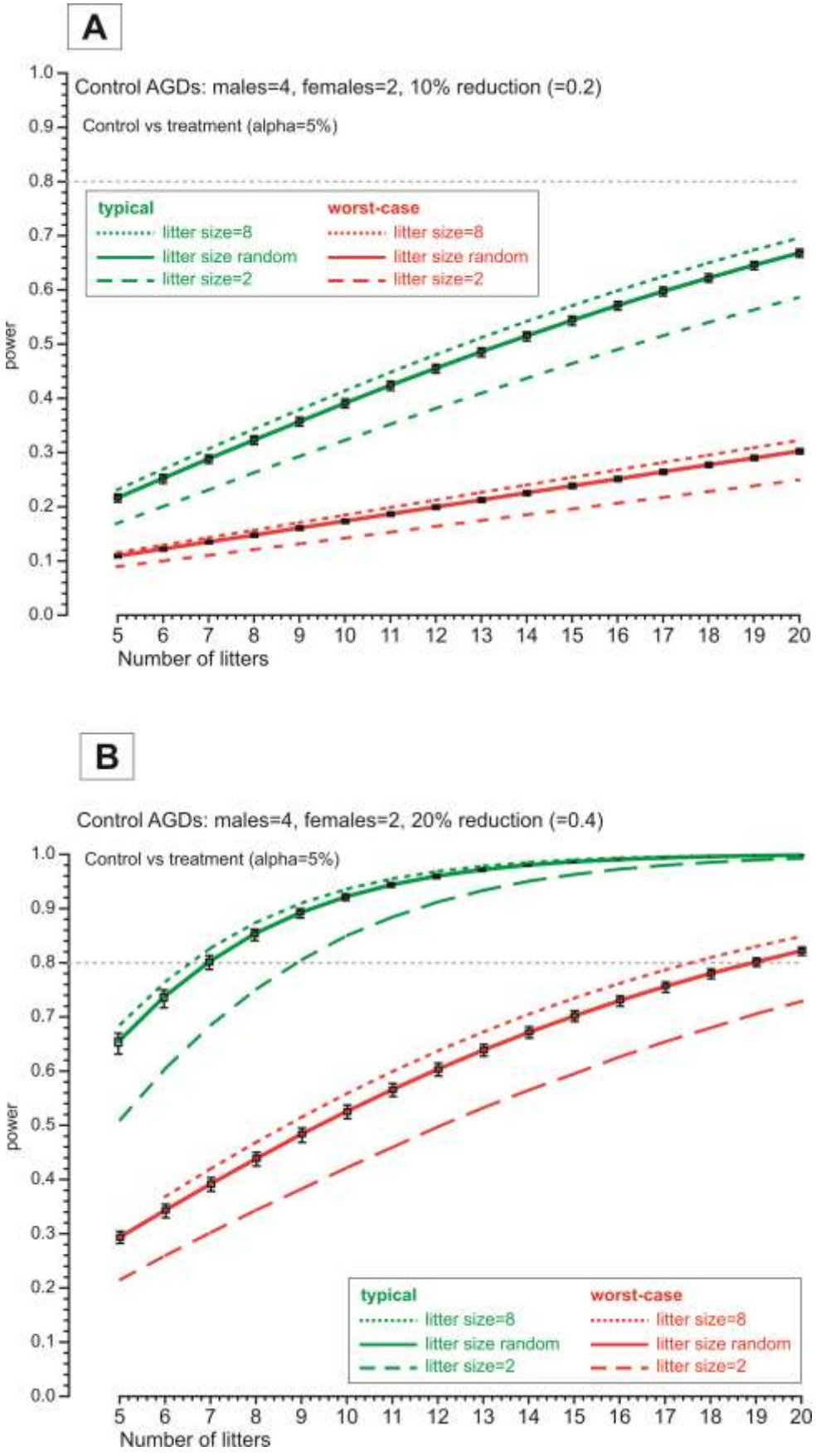
All simulations on the basis of these data sets indicate that the detection of effect sizes at same error rates as in the previous simulations requires much higher sample sizes, and a 10% reduction is likely to be overlooked by litter sizes below 20.

Table 2: AGD – Dose-Response summary (Non-Copenhagen studies)

Study	type of treatment	dose	litter #	litter av. size	birth weight mean	birth weight 90% percentile	Model information (LMM) between-litter variance	Model information (LMM) within-litter variance	Model information (LMM) ILC	Model information (LMM) tet_{BW}	AGD (mm) mean	AGD (mm) SEM	Control difference abs	Control difference ratio	Control difference norm	post-hoc power
A	male control		27	7.1	7.30	[6.60-8.20]					3.91	0.052	0	1	1	-
	Compound	NOAEL[1)]	24	7.8	7.05	[6.10-8.10]	0.051	0.101	0.34	1.89**	3.74	0.052	-0.17	0.96	0.91	32.4%
	female control		27	6.1	6.96	[6.00-7.80]					2.08	0.049	-1.83	0.53	0	-
B	male control		26	6.5	7.31	[6.30-8.30]					3.88	0.078	0	1	1	-
	Compound	NOAEL[1)]	26	6.3	7.20	[6.20-8.00]	0.155	0.142	0.52	1.76**	3.87	0.083	-0.01	1.00	0.99	17.5%
	female control		26	6.4	6.86	[5.90-7.80]					2.08	0.053	-1.80	0.54	0	-
C	male control		26	5.4	6.11	[5.50-6.85]					3.48	0.056	0	1	1	-
	Compound	NOAEL	30	5.3	6.27	[5.80-6.80]	0.076	0.059	0.56	1.37**	3.55	0.056	0.07	1.02	1.05	9.8%
		LOAEL	28	5.3	6.01	[5.20-6.70]					3.68	0.055	0.20	1.06	1.14	74.7%
	female control		26	5.5	5.76	[5.10-6.60]					2.03	0.016	-1.44	0.58	0	-
D	male control		28	5.5	6.25	[5.60-7.00]					3.80	0.027	0	1	1	-
	compound	NOAEL	27	5.9	6.15	[5.40-7.00]	0.030	0.066	0.31	1.05**	3.78	0.029	-0.02	1.00	0.99	40.4%
		LOAEL	29	5.6	5.94	[5.40-6.70]					3.64	0.052	-0.16	0.96	0.91	66.7%
	female control		28	4.9	5.98	[5.30-6.70]					2.06	0.025	-1.74	0.54	0	-
E	male control		27	5.0	6.12	[5.30-6.90]					3.43	0.041	0	1	1	-
	compound	NOAEL[1)]	28	4.8	5.88	[5.20-6.50]	0.038	0.058	0.39	1.57**	3.47	0.044	0.04	1.01	1.03	51.4%
	female control		27	5.6	5.79	[5.20-6.40]					1.68	0.033	-1.75	0.49	0	-
F	male control		27	4.9	6.21	[5.60-6.90]					4.13	0.051	0	1	1	-
	Compound	NOAEL[1)]	30	4.2	6.22	[5.50-6.90]	0.041	0.060	0.40	1.32**	4.21	0.041	0.09	1.02	1.04	24.5%
	female control		27	5.4	5.85	[5.20-6.30]					2.11	0.037	-2.01	0.51	0	-
G	male control		10	6.8	10.39	[8.90-11.90]					4.62	0.103	0	1	1	-
	female control		10	5.9	9.60	[8.60-10.90]					2.28	0.049	-2.33	0.49	0	-

ILC=inter-litter correlation; * stat. significant at α=5%; ** stat. significant at α=1%;
[1)] all doses produced significant responses, values for the lowest dose are shown; [2)] all doses produced non-significant responses, values for the highest dose are shown;
[3)] only two treatment doses were tested; tet_{BW} model parameter estimated for cubic root-transformed body weight;

Figure 7: Power simulation for AGD in male pups. 10% reduction corresponds to a 5% reduction in relation to the control male AGD. 20% reduction corresponds to a 10% reduction



Nipple Retention

Description of Data

Data sets from 20 studies were analysed, the summarizing information is shown in Table 3. The number of litters and average litter sizes were nearly identical to that reported in Table 1 for AGD. If data variability was small, average nipple numbers of below 1 could be detected as statistically significantly different from the controls. Litter correlations ranged from 0 to 0.5, indicating a weak to moderate intra-litter variation.

Description of Statistical Model and Estimation Method

The analysis of correlated data when the measurements are assumes to be multivariate normal has been studied extensively. However, when the responses are discrete and correlated, as this is the case for the number of nipples, different methodologies must be used in the analysis of data. As a general and flexible method for correlated discrete data, the generalized estimating equations (GEE) approach has become increasingly important and widely used in analysing such data (Liang & Zeger, 1986). In addition to those with the GEE approach there are other estimation approaches (e.g., weighted least squares), but here we only consider the GEE approach due to its increasing use in the field. The number of nipples/areolas was assumed to follow a binomial distribution with a response range between 0 and 12, with the latter being equal to the biologically possible maximal number of nipples in rats.

An attractive property of GEE is that working correlation matrix can be miss-specified and yet the regression coefficient estimator is still consistent and asymptotically normal. The covariance matrix of the estimated regression coefficients was estimated using the so called robust or sandwich estimator, the working correlation structure was chosen according to the working independence model. The correlated data can then be treated as though they were independent and the resulting regression parameter estimates along with the robust covariance estimator can be used to draw proper statistical conclusions. In general there is no closed form available which would allow sample size and power calculation in GEE (except for some special cases, but not relevant here). Note that one needs to specify the underlying correlation structure in sample size and power calculations and thus can use it as the working correlation structure.

Main assumption is that the pups from the same litter are correlated while litters are independent, i.e. within-litter correlation is present, but observations from different litters are independent. The number of nipples per pup was modelled by the marginal logistic regression model, where the unknown regression coefficients corresponded to the control and treatment mean estimates. We assumed that the correlation structure does not change across the litters, i.e. all pups had the same treatment-independent correlation (exchangeable compound symmetry matrix). The correlation parameter is defined between -1 and 1, with 1 assuming full correlation (all pups from the same litter responded in the same way), 0 no correlation at

all, and negative values indicate negative relationships. The latter was ruled out from a biological point of view, and consequently the correlation parameter was assumed to be positive, i.e. pups from the same litter are likely to respond more similar. This property can be translated as justification of using the litter mean in statistical analysis when the correlation parameter is close to 1, and using the pup response as statistical unit when the parameter is close to 0. Values between these two extremes therefore provide an estimate about the importance of the factor litter in data analysis, and its setting is crucial for sample size and power calculation. The between-litter variability ("variance estimate") was derived from the inference of the mean estimates, and used to define the variability scenarios for the power simulation studies.

Non-variation in the controls is problematic for statistics, i.e. if for none of the pups a nipple was measured. In a strict sense no data analysis can be done then. To overcome this limitation, a pragmatic solution might be setting a positive nipple for at least one pup per treatment group. However, for none of the selected data sets this was required.

All data analyses were performed by using the GENMOD procedure in the statistical software SAS.

Description of Simulations

For the Monte Carlo approach it was necessary to develop a method to simulate high-dimensional correlated data, in our case a high-dimensional multivariate binary distribution that describes the whole litter. This was achieved by using the copula techniques which combines marginal distributions with a given correlation structure (for more details see R. Wicklin, 2013). In contrast to the power simulation studies for AGD no approximate test statistics were available that could have simplified the simulation studies, and consequently the computer time required for the simulation studies was immense. This meant that it was not possible to estimate the detection limit at given error rates, only figures showing the power for specific experimental setups and data variability assumptions were produced. In each simulation step litter size was resampled by random out of a pool of all observed litter control sizes. All data analyses were performed by using the IML procedure in the statistical software SAS.

Results

The outcomes of the power simulations are shown for male controls with a small nipple baseline (0.1 nipples per pup) for two different variability scenarios in Figure 8, assuming the detection of 1 nipple per pup as effect size of interest. Here 10 litters per dose should be sufficient to ensure the statistical detection of this effect size. For the higher base line of 2 litters per control pup the detection of effect differences becomes much more difficult, and an increased nipple variability allows is likely to detect only large effect differences (Figure 9):

the detection of 4 nipples per pup at high data variability is unlikely to be achieved with litter numbers below 20. The negative impact of increased control baseline rates on statistical power is well-known from cancer studies with dichotomous tumour endpoints.

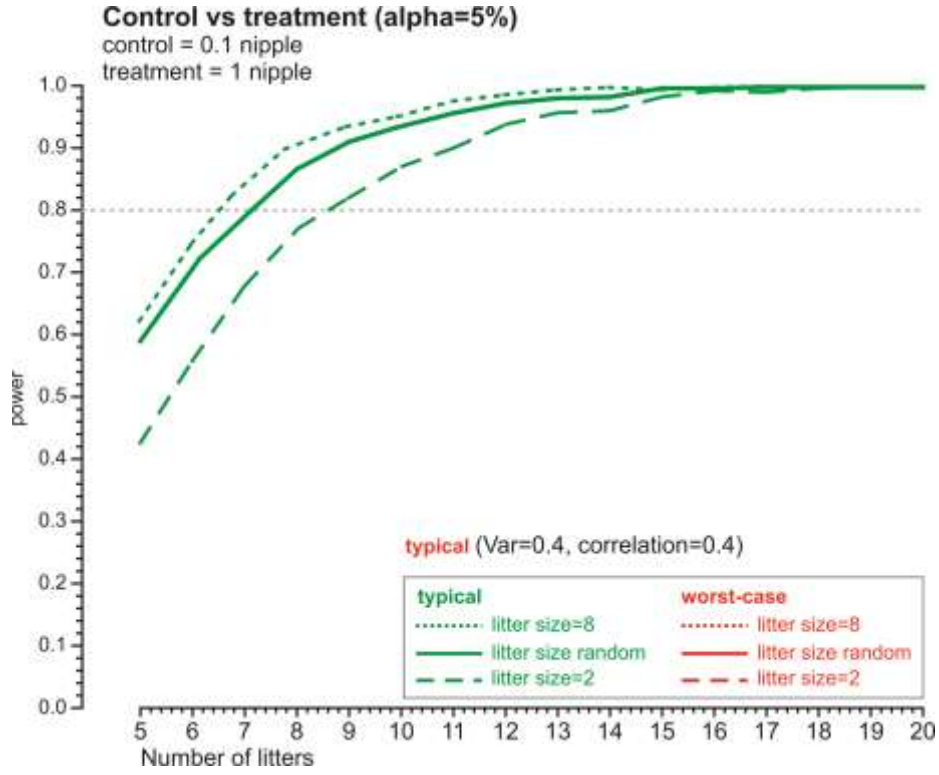
Table 3: Nipple Retention Summary for all Copenhagen studies

			litter			model information				
Study	**type of treatment**	**dose**	#	average size	mean nipple	logit link	Variance estimate	Litter correlation	Control difference	post-hoc power
A	male control		7	5.4	2.31	-1.533	0.0150			
	compound A	NOAEL	8	4.4	1.95	-1.736	0.0161	0.056	-0.36	22.3%
		LOAEL	10	5.1	2.85*	-1.268	0.0097	0.056	0.54	53.0%
B	male control		13	3.8	0.02	-6.484	0.9219			
	compound B	LOAEL[2)]	5	5.8	0.89*	-2.611	0.1408	0.331	0.87	57.4%
	compound C	LOAEL[2)]	7	3.4	1.28*	-2.215	0.1232	0.367	1.26	60.9%
C	male control		15	4.1	0.23	-4.030	0.1334			
	compound D	NOAEL	8	4.0	0.89	-2.615	0.3885	0.350	0.66	13.8%
		LOAEL	8	3.8	3.95**	-0.829	0.1047	0.350	3.72	96.7%
	compound E	LOAEL[2)]	8	3.6	3.78**	-0.890	0.1052	0.332	3.55	98.0%
D	male control		15	4.3	0.40	-3.461	0.4052			
	compound F	NOAEL[1)]	6	5.2	2.66	-1.356	0.0826	0.506	2.26	15.3%
	compound G	NOAEL[1)]	8	4.0	0.89	-2.608	0.2413	0.409	0.49	5.1%
	compound H	NOAEL	8	5.3	1.38	-2.127	0.0532	0.248	0.98	9.5%
		LOAEL	5	4.6	3.72*	-0.913	0.1291	0.248	3.32	50.7%
E	male control		13	4.8	0.02	-6.612	0.9072			
	Mixture A	LOAEL[2)]	14	4.4	1.01**	-2.478	0.0623	0.140	0.99	99.9%
F	male control		13	3.8	1.60	-1.965	0.0847			
	compound I	NOAEL[1)]	7	3.0	3.61	-0.942	0.0915	0.212	2.01	29.3%
	compound J	NOAEL[1)]	7	4.3	3.08	-1.168	0.0869	0.462	1.48	1.5%
	Mixture B	LOAEL[2)]	16	4.8	3.65*	-0.950	0.0554	0.385	2.05	95.2%
G	male control		13	4.3	0.06	-5.333	0.1341			
	Mixture C [3)]	NOAEL[1)]	14	5.9	0.53**	-3.151	0.0710	0.128	0.47	93.1%
H	male control		19	4.4	0.01	-6.997	0.9399			
	Mixture D	NOAEL[1)]	19	4.6	0.69	-2.877	0.1187	0.104	0.68	44.3%
	Mixture E [3)]	LOAEL[2)]	15	4.6	1.13*	-2.352	0.0447	0.048	1.12	71.4%
	Mixture F [3)]	NOAEL[1)]	17	5.2	0.04	-5.667	0.2497	<0.01	0.03	2.3%
I	male control		18	6.3	0.09	-5.011	0.1682			
	compound K	NOAEL[2)]	17	6.3	0.43	-3.371	0.1226	0.121	0.34	
K	male control		15	5.1	0.14	-4.553	0.1798			
	Mixture G [3)]	LOAEL[2)]	16	5.6	0.66**	-2.926	0.0711	0.173	0.52	94.9%
	Mixture H [3)]	NOAEL[1)]	16	4.9	0.14	-4.492	0.1312	0.034	0.00	0.5%

Mixture I [3)]	LOAEL[2)]	17	5.8	1.55**	-1.997	0.0268	0.139	1.41	99.9%.

* stat. significant at α=5%; ** stat. significant at α=1%; logit link = mean model estimate after logit transformation; [1)] all doses produced significant responses, values for the lowest dose are shown; [2)] all doses produced non-significant responses, values for the highest dose are shown; [3)] only two treatment doses were tested;

Figure 8: Power simulation for nipple retention in male pups with low control baseline.



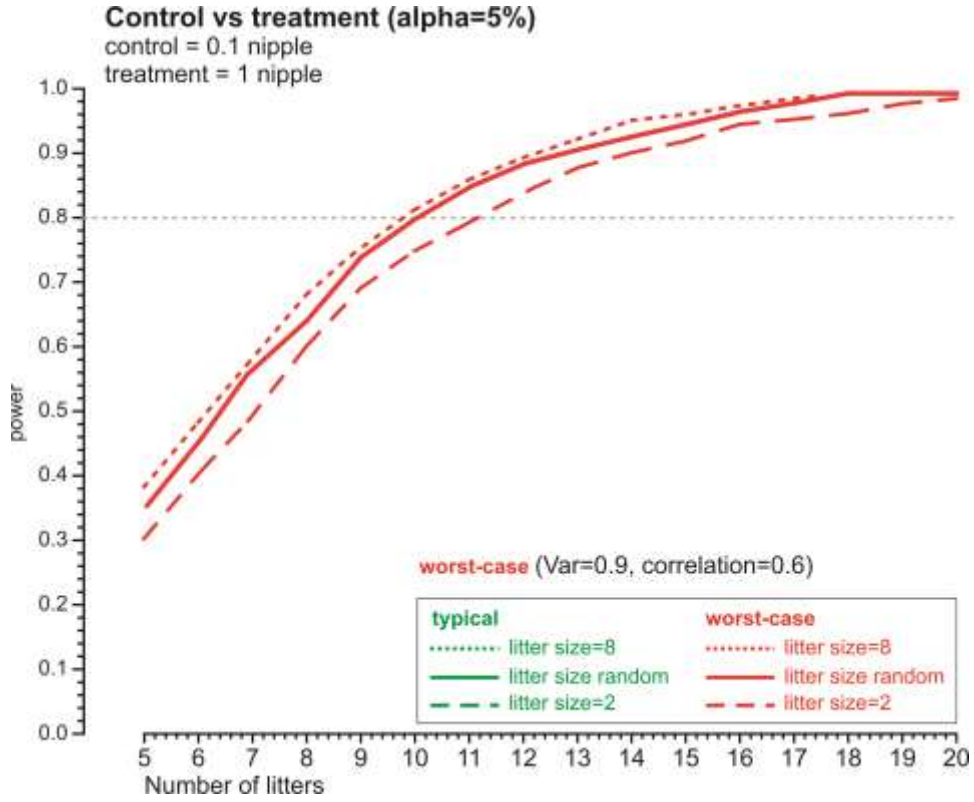
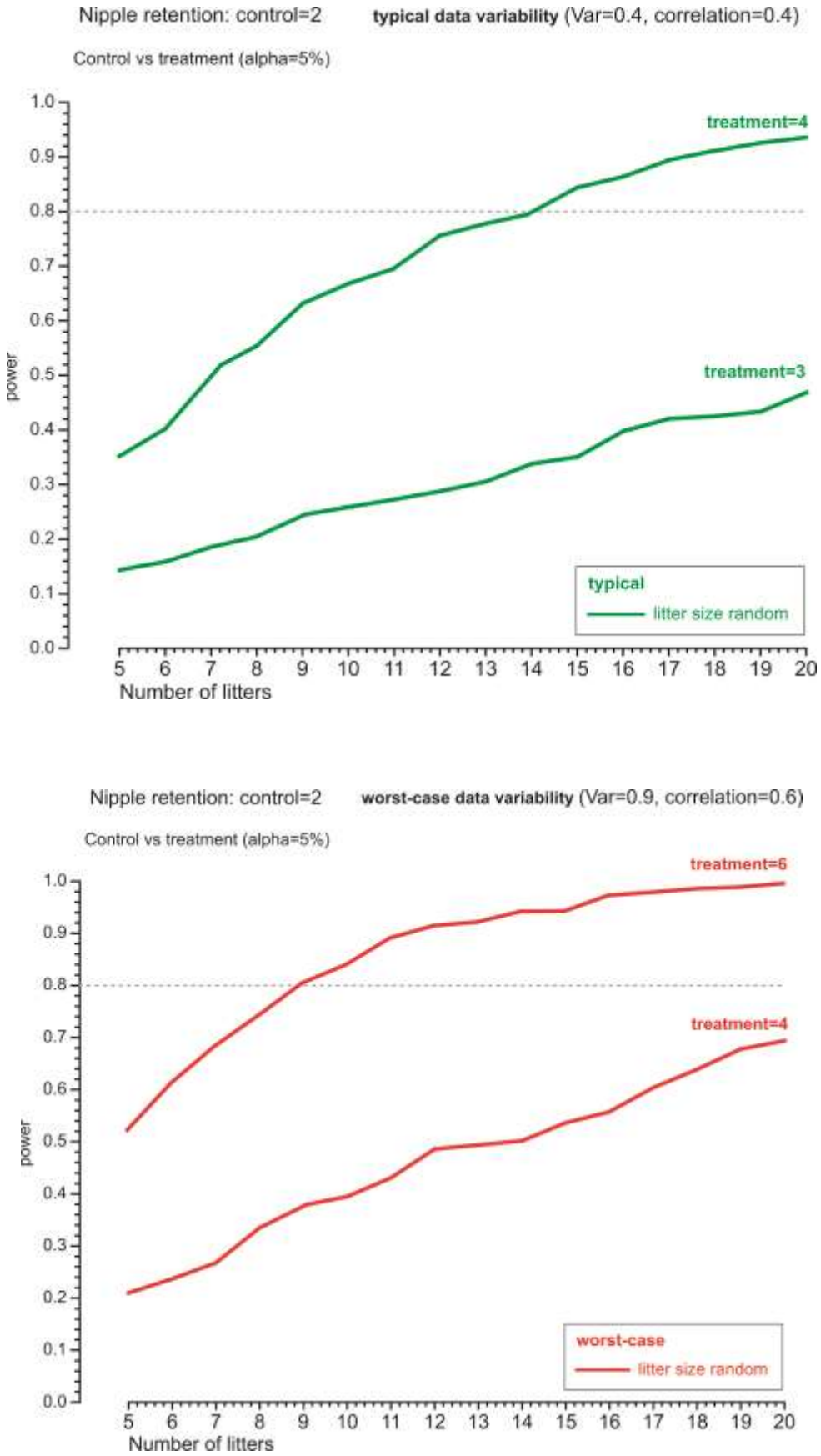


Figure 9: Power simulation for nipple retention in male pups with high control baseline.



Comparison AGD vs. NR

The successful NOAEL and LOAEL determinations for all studies are shown in Table 4. If only a NOAEL was derived, then none of the tested doses produced responses significantly different from the controls, and if only a LOAEL was determined, then all of the tested doses produced statistically significant responses. Also included is whether the NOAELs agreed for both endpoints, or whether different conclusions were drawn.

Table 4: NOAELs & LOAELS for AGD and NR and their comparative assessment

		AGD		NR		comparative assessment ("sensitivity")		
Study	**type of treatment**	NOAEL	LOAEL	NOAEL	LOAEL	NR < AGD	NR = AGD	AGD < NR
Copenhagen studies								
A	compound A	X	-	X	X	X		
B	compound B	X	X	-	X	X		
	compound C	-	X	-	X		X	
C	compound D	X	X	X	X	X		
	compound E	-	X	-	X		X	
D	compound F	X	-	X	-		X	
	compound G	X	-	X	-		X	
	compound H	X	X	X	X		X	
E	Mixture A	X	X	-	X	X		
F	compound I	X	-	X	-		X	
	compound J	X	-	X	-		X	
	Mixture B	-	X	-	X		X	
G	Mixture C	X	-	-	X	X		
H	Mixture D	X	X	X	-			X
	Mixture E	-	X	-	X		X	
	Mixture F	X	-	X	-		X	
I	compound K	-	X	X	-			X
J	compound L	X	X	X	-	X		
K	Mixture G	-	X	-	X		X	
	Mixture H	X	-	X	-		X	
	Mixture I	-	X	-	X		X	

References

Bretz F and Hothorn LA (2003) “Statistical Analysis of Monotone or Non-monotone Dose Response Data from In Vitro Toxicological Assays”. ATLA 31:81-96

Liang KY and Zeger SL (1986) “Longitudinal Data Analysis Using Generalized Linear Models,” Biometrika, 73, 13–22

Littell RC, Milliken GA, Stroup WW, Wolfinger RD, and Schabenberger O (2006) *SAS for Mixed Models*, Second Edition, Cary, NC: SAS Press

McCullagh P and Nelder JA (1989) *Generalized Linear Models*, Second Edition, London: Chapman & Hall.

Stroup WW (1999) “Mixed Model Procedures to Assess Power, Precision, and Sample Size in the Design of Experiments,” 1999 Proceedings of the Biopharmaceutical Section, Alexandria, VA: American Statistical Association, pp. 15–24

Verbeke G and Molenberghs G (2000) *Linear Mixed Models for Longitudinal Data*, New York: Springer

Wicklin R (2013) *Simulating Data with SAS*, Cary, NC: SAS Institute Inc

APPENDIX 1

POWER SIMULATIONS OF T4 HORMONE LEVELS OF RODENTS EXPOSED TO ENDOCRINE DISRUPTING CHEMICALS

Objective

This power study is an extension to the report *Power Simulations of Nipple Retention and Anogenital Distance of Rodents exposed to Endocrine Disrupting Chemicals (Annex 1a)*, with the objective to determine the optimal allocation of litter numbers per dose in order to study the effect of certain endocrine disrupting chemicals on Thyroid hormone blood levels (T4) measured in pups and dams. This analysis was performed considering that the blood samples from the pups killed are pooled for males and females from the same litter (i.e. one measure per litter), and compared to data sets of gender-specific non-pooled measurements. The conclusions reached are summarized next, followed by a detailed justification for these conclusions. Details about statistical testing and the problematic adjustments to p-values for multiplicity, statistical error rates, the power concept and NOAEL determination can be found in the first report and are here not described. It is assumed that all statistical analysis is based on data obtained under comparable testing conditions, as outlined in the main report.

Conclusions:

1. Conclusions about intra-litter variability could not be drawn, and thus no statement about the statistical uncertainty of a litter mean. As consequence it is unclear what the optimal allocation of pups for the estimation of a litter mean is. The litter mean is considered as an independent non-random variable in data analysis.

2. Data from DTU revealed in average the smallest variability between T4 litter means. The reasons for the differences in litter mean variations between DTU and the other labs are unknown.

3. Variability between litter means can be huge (CV>60%), and it is unclear whether this is caused by biological or technical factors.

4. T4 litter responses are more homogenous from controls and older pups.

5. Variability is likely to be higher at high effect doses.

6. Pooling T4 levels from pups from the same litter provided no sound indications for a more robust litter mean estimation if compared to using only one pup per litter.

7. In none of the studies gender-specific differences in the T4 levels were detected in the control animals, but two studies revealed under dosing statistically significant different T4 responses between the sexes.

8. The detection of T4 changes by 20% might be too optimistic, as it assumes a rather low to moderate data variability for this endpoint (CV <= 20%). In worst case (CV=60%), even with high litter numbers (N<30) a 50% T4 change might be undetected.

Hint: All statistical analyses were done on the basis of the coefficient of variation (CV).

Data availability

The majority of data were provided from DTU (Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark) and US EPA, remaining data sets were made available from two member of the OECD expert group.

Data from the US EPA was based on T4 levels derived from pooled tissue samples from an equal number of male and female pups from the same litter, and measured at different developmental stages of the pups. In addition T4 levels were measured in dams (at birth and PND21). Litter mean was considered always as the statistical unit, and data summaries were provided for each control or treatment groups, expressed as litter means plus standard deviation (or standard error of the mean) and corresponding litter number. This suggests that data have been tested for normality. Litter number per control or dose group varied between 5 and 26. All data have been published.

Data provided from DTU were from three different studies (Wistar rat), where in two studies T4 levels were measured from one pup per litter and for both sexes, at different pup ages, and from the dams, otherwise T4 levels were pooled from male and female pups from the same litter. The number of litters in the investigated dose groups varied between 5 and 20. A different method for measurement of T4 levels was used in the first DTU study, compared to the last two. None of these data have been published.

Remaining data sets were from three studies, with measurements from the pups at PND4 (test species not reported), and litter means determined on the basis of pooling or individuals. Both studies included 4 treatment doses, with data available from 8-20 litter per group.

Data means with reported zero standard deviation or CVs above 100% were excluded from all data analysis. The latter might indicate the overlooked presence of an outlier in the rough data. Samples with less than 5 litters were also excluded from data analysis. It should be noted that at high effect doses occasionally levels below the limitation of quantification (LOQ) were measured, which might have violated the ANOVA assumptions and thus resulted into unreliable CV estimations. Furthermore, control levels in pups from GD0-GD4 are generally lower than from later stages, and therefore reducing the range of possible T4 reductions.

Data treatment

T4 levels were measured at different developmental stages in pups and dams and from different rodent species, and data were provided in different units (ng/ml, nM). To achieve a better data comparability, all statistical analysis were done on the basis of the coefficient of

variation (CV), which is defined as the ratio of the standard deviation to the mean of the sample. However, as most sample sizes can be considered from a statistical point of view as rather small, this calculation leads to biased estimator. Therefore the CV calculation was corrected as

$$CV_{corrected} = \left(1 + \frac{1}{4 * N}\right) * CV,$$

with N sample size (Sokal RR & Rohlf FJ, 1995). As consequence, the lower the sample size, the larger the correction, e.g. at N=10 the CV is increased by a factor of 2.5%. In the following CV refers always to this equation. T4 measurements were always normalized to the control mean, and only these values were used in power analysis, i.e. responses are expressed as relative to the control mean.

Data description and analysis followed always the same purpose: identifying a reference measure of data variability for the power and sample size analysis. As no intra-litter information was available (i.e. more than one T4 measurement per litter), the CV always expresses the overall between-litter variability within a control or dose group. This simplified not only the power analysis as only one source for variation had to be considered, the use of the CV also meant that the absolute scale of the mean T4 estimate was not relevant for the data analysis.

Prior the power calculations the following questions were investigated:

- Do the data provide indications for gender differences in the T4 levels?
- By pooling the serum levels of pups from the same litter, can we expect a smaller variability between litter means? (as indication for a more robust litter mean estimate)
- Which factors influence the litter variability?

Statistical Testing

As no indications against the assumption of Gaussian distributed effect data were found, Analysis of variance (ANOVA) was used to analyze for effects of treatment. Due to the lack of homogeneity of variance, the Behrens-Fisher Students' t-test was used always with a Satterthwaite approximation for the degrees of freedom.

Data description

Each individual CV describes the scatter of T4 responses between litters from the same control or dose group. The variation of CVs across different studies and dose groups and for different age classes is summarized in Table 1. In order to identify CV changes in relation to the T4 responses, dose groups were divided into a “low effect” group with mean T4 levels above 80% of the controls, and a “high” group with mean T4 levels below 80% of the

controls. N refers always to the number of data sets, in case of the controls it equals the number of studies.

As T4 measurements were available only from pups at PND4, the remaining data sets are not shown. Here the CVs ranged from 16% to 60%.

Table 1 shows huge differences between the CVs, in worst cases litter variations with CVs above 60% were estimated. Also significant differences between the labs were observed: although less data with T4 measurements were provided by DTU, it provides clear indications that in average their data variability is much smaller than from other labs.

Table 1: Coefficient of variations (CV) for T4 litter samples from dams and pups of various age classes.

			Dam		Pup					
			GD15	at birth	PND0-1	PND3-4	PND7-8	PND14-16	PND21-22	PND27
DTU	Gender									
Control	male	N	-	-	-	-	1	3	-	1
		CV_{range}	-	-	-	-	19.4	6.4 [4.3;15.5]	-	16.3
	female	N	2	-	-	-	1	3	-	1
		CV_{range}	12.0;12.5	-	-	-	27	8.9 [7.5;18.5]	-	14
	pooled	N	-	-	-	1	1	-	-	-
		CV_{range}	-	-	-	8.2	9.14	-	-	-
Low[1)]	male	N	-	-	-	-	-	-	-	-
		CV_{range}	-	-	-	-	-	-	-	-
	female	N	1	-	-	-	-	-	-	-
		CV_{range}	12.9	-	-	-	-	-	-	-
	pooled	N	-	-		3	-	-		
		CV_{range}	-	-		23.0 [21.7;34.6]	-	-		
High[2)]	male	N	-	-	-	-	4	9	-	2
		CV_{range}	-	-	-	-	18.9 [15.9;23.1]	17.4 [12.2;26.3]	-	16.8;19.6
	female	N	4	-	-	-	4	9	-	2
		CV_{range}	14.0 [9.8;20.0]	-	-	-	20.6 [7.7;22.9]	12.2 [3.6;38.9]	-	10.0;13.8
	pooled	N	-	-	-	3	3	-	-	-
		CV_{range}	-	-	-	6.1 [4.2;7.8]	7.2 [6.8;8.9]	-	-	-
US-EPA			**at birth**	**PND21**	**PND0-1**	**PND3-4**	**PND7-8**	**PND14-16**	**PND21-22**	**>PND36**
Control	pooled	N	4	8	3	7	3	11	9	2
		CV_{range}	32.3 [21.4;38.7]	18.2 [14.0;26.6]	18.5 [15.6;44.8]	31.4 [15.3;76.6]	14.1 [12.7;29.6]	13.3 [11.7;33.1]	18.1 [8.9;46.1]	17.9;21.1
Low[1)]	pooled	N	3	12	2	11	1	12	10	6
		CV_{range}	44.2 [15.4;68.1]	26.9 [14.2;58.1]	13.6 [9.7;17.6]	29.7 [18.1;66.6]	38.4	19.6 [10.1;39.9]	16.1 [10.6;33.5]	15.2 [11.7;18]

High[2)]	pooled	N	10	15	4	10	4	19	14	-
		CV_{range}	37.8 [3.5;87.4]	30.8 [20.0;71.6]	40.8 [9.7;52.6]	39.8 [25.6;80.0]	33.3 [26.3;45.3]	42.8 [0.1;89.5]	38.4 [14.0;62.8]	-

N= number of groups, CV_{range} = median with minimum and maximum (in brackets)
[1)] samples from low effect doses (T4 levels >80%); [2)] samples from high effect doses (T4 levels <=80%);

Gender differences

Two labs provided measurements from individual pups which allowed the investigation of possible gender differences. In these studies T4 levels from only one pup per litter were analyzed, and as consequence it was possible to test for gender differences only across litters, but not within the same litter. Data from four studies were available, with three providing information from more than one developmental age. None of the control groups provided any indications for a general gender difference (N=6), neither in terms of statistical significance nor of a higher preference of higher (or smaller) values for one gender. However, under dosing two samples (out of N=18) yielded different responses: in one study (DTU), two doses produced T4 concentrations that were stat. significant higher in male pups (PND27), in the other study (data from a different lab, PND4) the opposite was observed, here females provided higher measurements at lowest treatment dose (not confirmed at higher doses in the same study). In none of the studies a data pooling would have produced different dose-response results when gender information is ignored. Whether these observed gender differences are true or only false-positives cannot be answered with confidence, and more empirical evidence is certainly required. Nevertheless, a gender effect cannot completely ruled out, and if present (however small), we would expect an increase in data variation within the litter if data from the two genders are pooled and used in data analysis. However, this does not mean necessarily that the between-litter variability will also increase. If always an equal number of male and female pups have been used for the litter mean estimation, which is the preferred balanced design approach by US EPA, and more than one pup per sex per litter is used, than the precision of the litter mean can be higher than based on only one pup.

To investigate whether the pooling of T4 values from male and female pups (although from different litters) could increase the data variability of T4 litter means, CVs were calculated for both genders individually and together and then compared. This is illustrated in Figure 1: on the x axis the CVs for both genders are shown, and on the y axis the corresponding value if both data sets are pooled. The black trend line indicates a perfect 1:1 relationship between both CVs, i.e. the case of no gender contribution to the CV, and the red line is the regression line together with its 95% CI. Here we included only T4 values from dose groups, i.e. controls were excluded from data analysis. The linear regression line agrees well with the black reference line, and thus it provides no indications that pooling data from both genders will increase the overall variability between litter means. Therefore the factor gender was ignored in the power analysis, i.e. we assume that it doesn't matter whether the litter mean is measured from blood samples from one or more individual pups.

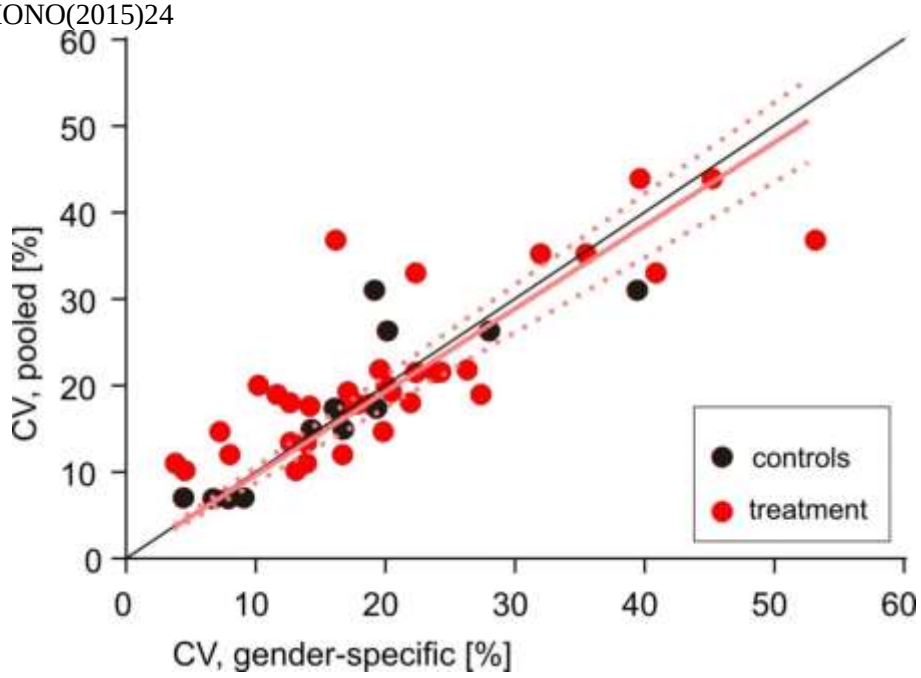


Figure 1 Variability of T4 measurements from male and female pups (x axis) vs. pooled (y axis, expressed by the Coefficient of variation (CV).

CV references for the power analysis

To identify the best CV candidates for the power analysis, data were further simplified with T4 levels from the controls and low effect doses grouped against those from the high effect dose, and litter means derived from individual pups were treated in the same way as pooled litter means. The results are shown in Table 2.

To provide more details about the distribution of CV values, data provided by the US EPA are shown in Figure 2, separated for dams and pups of different age classes. Each individual CV refers to a treatment group (control or exposure), derived from 5 to 28 litters. It should be noted that a general trend between litter size and the CV was not identified, i.e. high CVs are not due to small sample sizes (however, the precision of an estimated CV decreases with increasing sample size). From the graph it can be conclude that

- data scatter of T4 measurements can vary hugely between different treatment groups and studies,
- it is lower in control and low effect groups than high exposure groups,
- and the older the pups, the more likely to measure homogenous T4 responses, at least in the control and low effect groups.

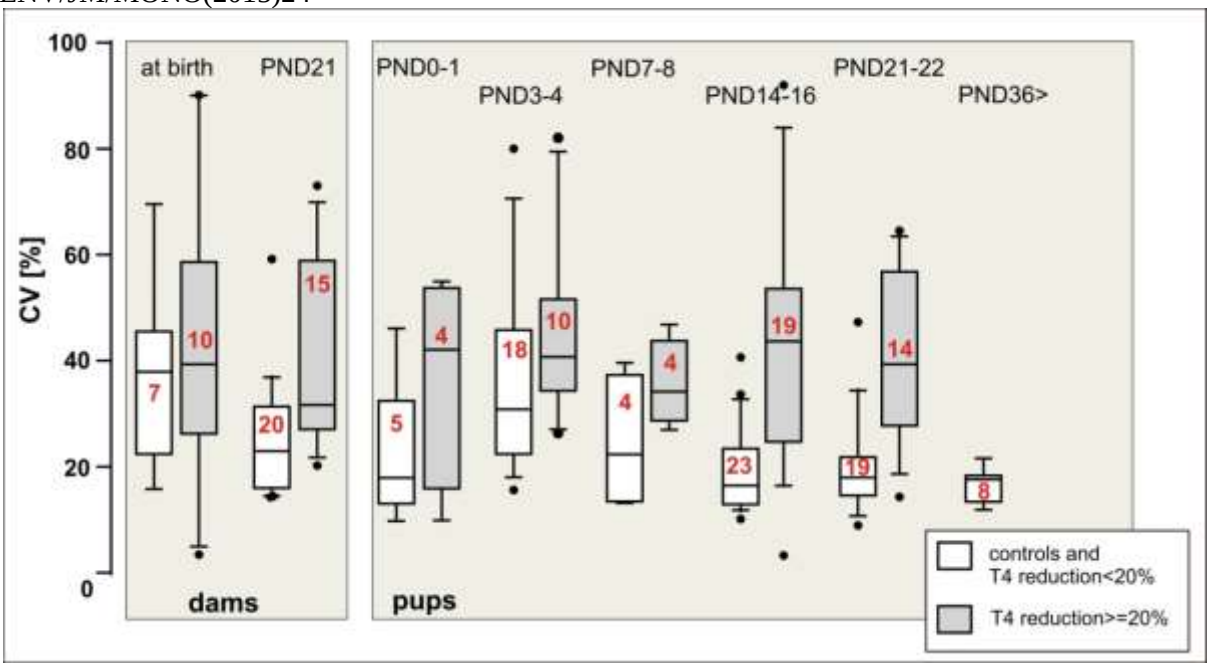
Altogether, a reference value well below 20% for the CV must be considered as a rather unlikely event. This might be achieved in some labs (see DTU), but not necessarily in all, so a CV=20% was regarded as an average expectation. For the worst-case scenario we used a CV=30%.

Table 2: Coefficient of variations (CV) for T4 litter samples from pups of various age classes.

DTU		Pup PND0-1	PND3-4	PND7-8	PND14-16	PND21-22	PND27
T4 reduction <20%	N	-	1	2	3	-	1
	CV_{median}	-	8.2	17.5	7.0	-	14.8
	$CV_{90\%}$ percentile	-	-	25.8	17.1	-	
T4 reduction >=20%	N	-	3	7	9	-	2
	CV_{median}	-	6.1	11.7	17.7	-	18.6
	$CV_{90\%}$ percentile	-	7.8	21.1	32.2	-	19.8
member data							
T4 reduction <20%	N	-	10	-	-	-	-
	CV_{median}	-	33.0	-	-	-	-
	$CV_{90\%}$ percentile	-	49.1	-	-	-	-
T4 reduction >=20%	N	-	2	-	-	-	-
	CV_{median}	-	27.6	-	-	-	-
	$CV_{90\%}$ percentile	-	29.2	-	-	-	-
US-EPA		**PND0-1**	**PND3-4**	**PND7-8**	**PND14-16**	**PND21-22**	**>PND36**
T4 reduction <20%	N	5	18	4	23	19	8
	CV_{median}	17.6	30.5	21.8	16.0	17.5	17.2
	$CV_{90\%}$ percentile	44.8	66.6	38.4	30.8	33.5	21.1
T4 reduction >=20%	N	4	10	4	19	14	-
	CV_{median}	40.8	39.8	33.3	42.8	38.4	-
	$CV_{90\%}$ percentile	52.6	67.4	45.3	81.7	60.4	-

N= number of data sets

Figure 2: Distribution of coefficient of variations (CV) from T4 litter samples. Only data from US EPA are shown. Numbers refer to the treatment groups, with controls and low effect exposure grouped together.


100
80
60
40
20
0
CV [%]
at birth
PND21
dams
7
10
20
15
PND0-1
PND3-4
PND7-8
PND14-16
PND21-22
PND36>
pups
5
4
18
10
4
4
23
19
19
14
8
controls and
T4 reduction<20%
T4 reduction>=20%

Power and sample size

As the mean T4 litter values fulfilled the ANOVA assumption of normality, a closed mathematical expression exists which determines the exact sample size at given power (or vice versa), and therefore it was not necessarily to perform computer-intensive simulation studies (DiSantostefano and Muller, 1995). The only requirement for the a priori power analysis and corresponding sample size calculation is knowledge about the expected variation of the litter means (here always expressed as CV), the T4 change of interest, and the false-positive rate we are willing to accept. The latter was set to α=5% (two-sided), and a false-negative rate of maximal β=20% (i.e. power of 80%) was considered as target. In a first step the p-value of the t-test was not adjusted for multiple comparison, meaning that the power analysis was performed for an individual comparison between control and treatment group. As there are various ways to control for the global false-positive rate (see first report), in the second step the p value was adjusted by a Bonferroni correction assuming an experimental setup of three treatment doses. The Bonferroni correction can be considered as the least powerful statistical correction method and is not recommended, as many more powerful approaches exists. Nevertheless, the range between the power curves derived from an unadjusted and Bonferroni corrected p value indicate the power which can be achieved with more advanced statistical approaches to control multiplicity. At given power, therefore both curves provide a range of sample sizes (i.e. number of litters) per group which is necessarily to identify the T4 change of interest at given error rates and data variability.

The detection of 20% or 30% changes in control T4 levels were considered of interest, and for each we used three different scenarios for the litter means: (i) the variability of the litter means in the control and dose group were set to CV=20%, corresponding to an average outcome of the previous data analysis (red line), (ii) the variability in the dose group were doubled (CV=40%), reflecting more the responses of a high effect dose (blue line), (iii) variability in both groups were set to CV=40% (worst-case data scenario) (green line). All data analyses were performed by using the statistical software SAS.

The results are shown in Figure 3 and Figure 4. If the detection a 20% change in the control T4 levels at high likelihood is of interest (power > 80%), then at least 17 litter per group are required, and if the data variation is above the average, then well above 30 litter are required (Figure 3). If the statistical detection limit is lowered to a 30% change, then a sample size of 10 litters is sufficient assuming average data variability (Figure 4). However, less homogenous litter responses will require much higher litter numbers even for this effect size. For example, with CVs of 60% the chance of detecting a 50% reduction in T4 levels as statistically significant is well below 50% for groups with up to 30 litters.

Figure 3: Power to identify a 20 % change in T4 levels at given litter number

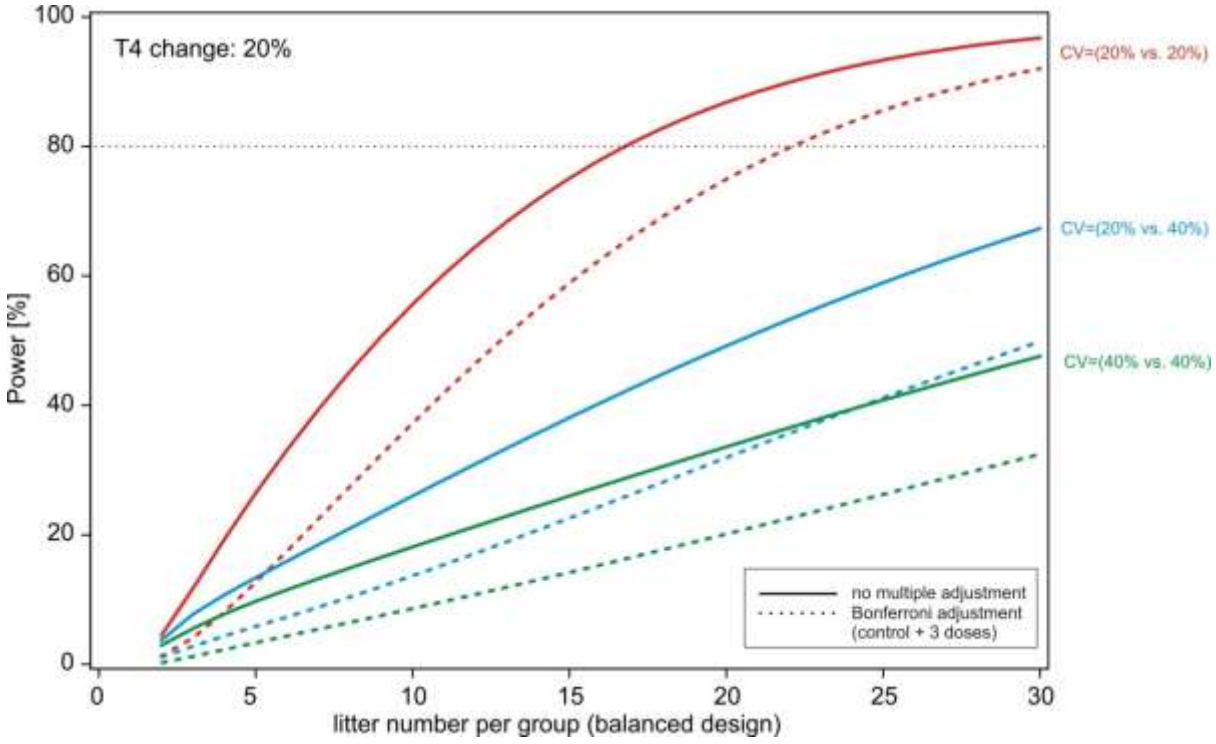
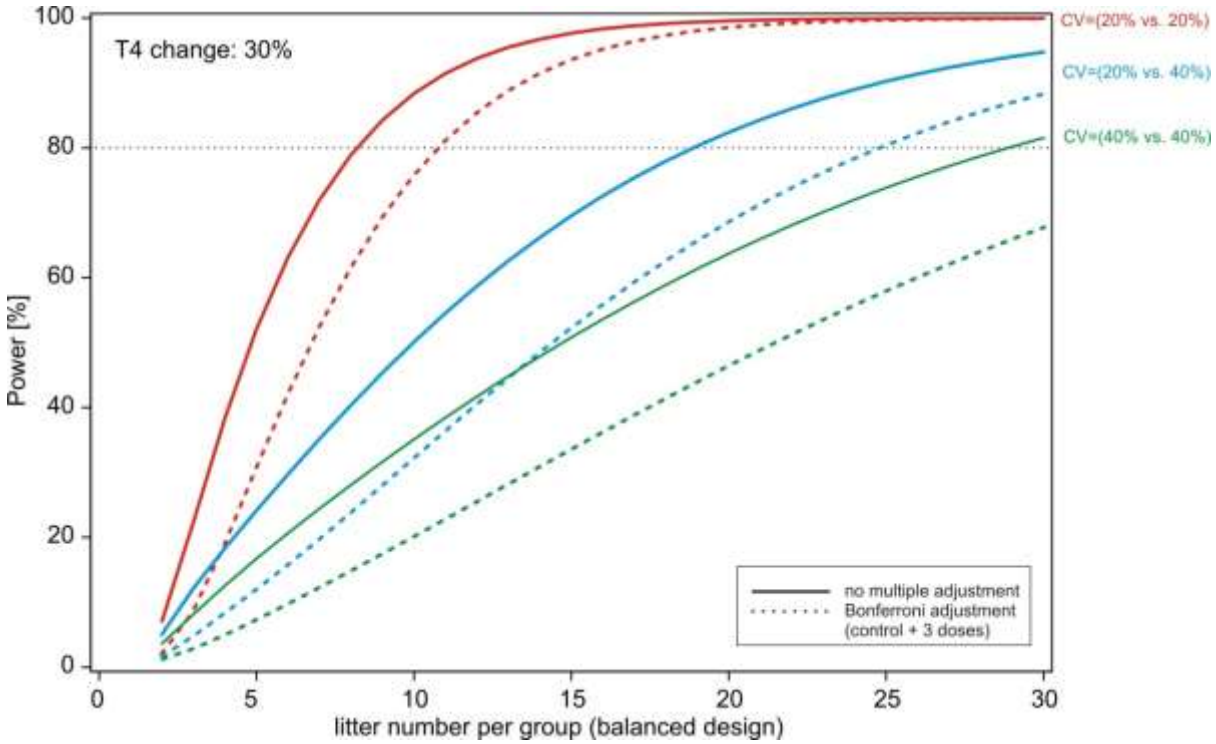


Figure 4: Power to identify a 30 % change in T4 levels at given litter number



References

DiSantostefano, R. L. and Muller, K. E. (1995), "A Comparison of Power Approximations for Satterthwaite's Test," Communications in Statistics—Simulation and Computation, 24, 583–593

Sokal RR & Rohlf FJ. Biometry (3rd Ed). New York: Freeman, 1995. p. 58

APPENDIX 2

STUDY REPORT FROM DTU FOOD ON MALFORMATIONS OF THE EXTERNAL GENITALIA IN YOUNG MALE RAT OFFSPRING

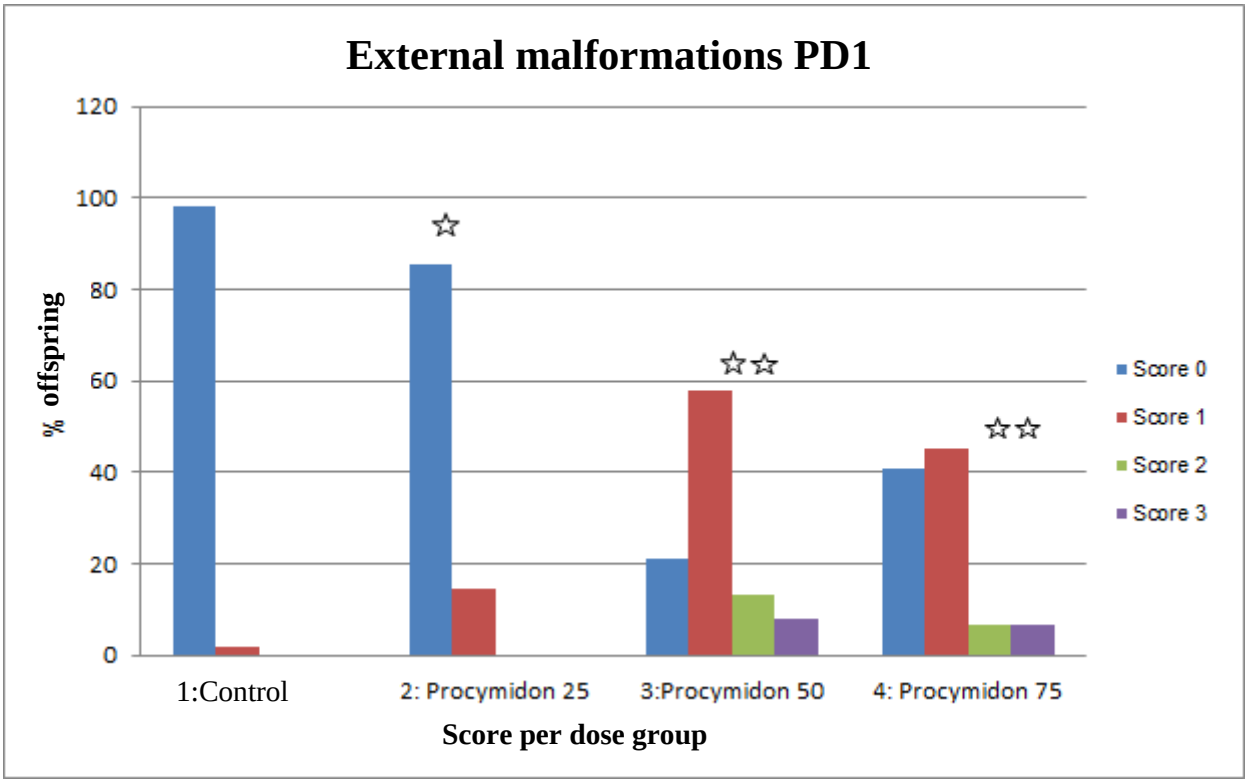
This project investigates sexual development in male rat offspring after *in utero* exposure to the endocrine disrupting anti-androgen procymidone. The main purpose of this study was to investigate whether malformations of the male offspring's genitalia could be scored soon after birth and furthermore to access weather it was possible to score the degree of these malformations early after birth. Moreover whether there is a correlation between anogenital distance (AGD) at birth, nipple retention at pup day (PD) 14 and malformations of the external genitalia in the young male rat offspring.

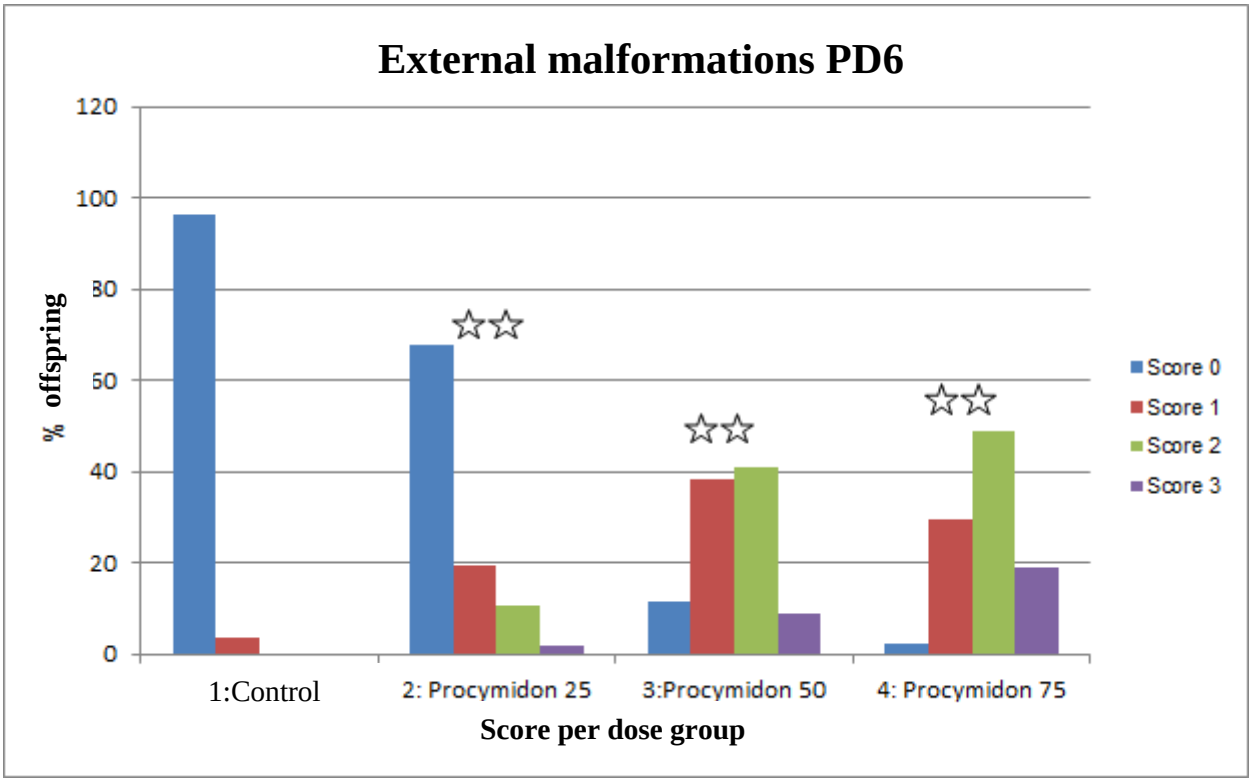
40 paired Wistar rats were dosed with either 0, 25, 50 or 75 mg / kg body weight / day of the known anti -androgenic pesticide procymidone from day 7 of gestation to PD 4. At birth the following endpoints was recorded: weight of the dams, sex ratio in litters, pup weight and AGD as well as external malformations of the sexual organs. A score of 0-3 was used for the external malformations. Score 0 indicated normal while score 3 indicated very severe malformations. All male rats were kept until PD 55. All male offspring were examined for external abnormalities of the genitals at PD 6, PD 14, PD 22 and PD 54 /55. In addition, the animals were tested for nipple retention at UD 14, and their penile length was measured at UD 55.

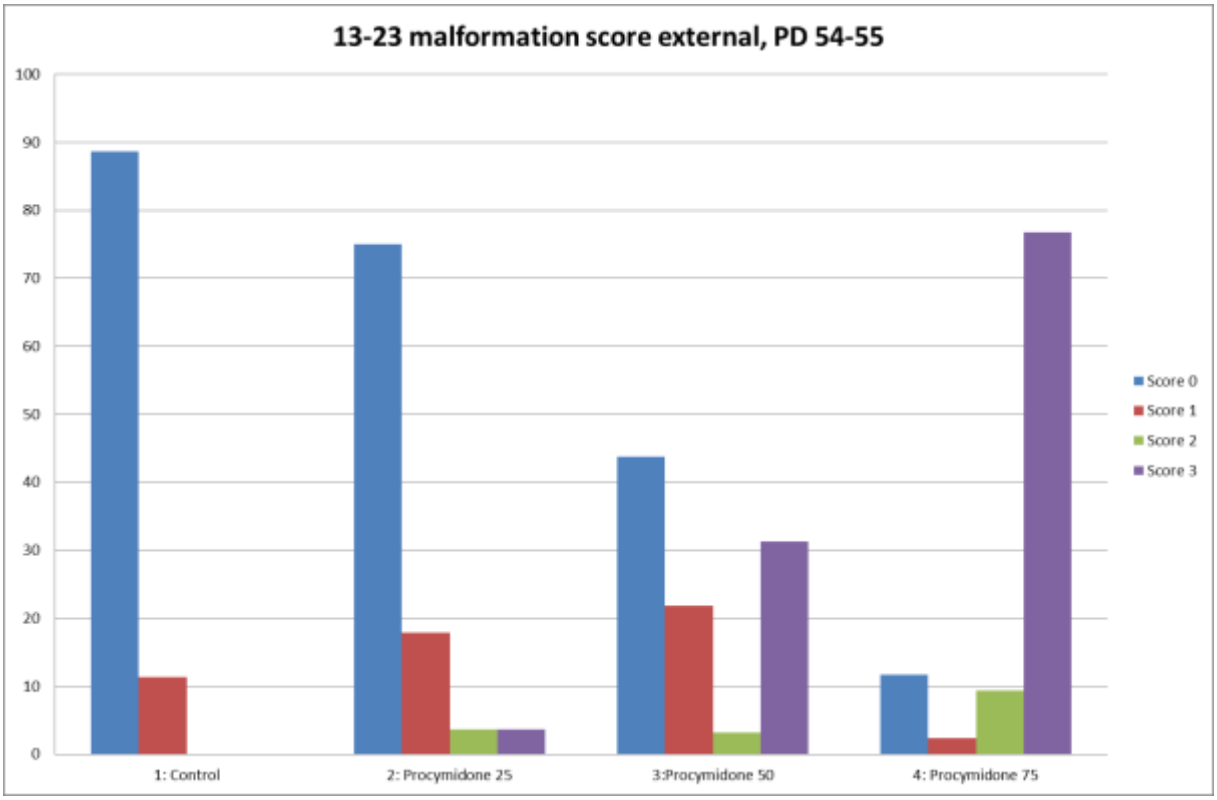
The study showed that AGD was significantly decreased in all exposed groups ($p < 0.01$) and that the males from all exposed groups showed a significant increase in the malformation score compared with the control males (at both UD 6, UD 14, UD 22 and UD 54 /55) . There was an increase in the number of males with malformations from day 0 to day 6, and a dose-related correlation between reduced AGD and the increased occurrence of malformations. This reduction was more evident on days 6, 14 and 22 than on day 0. On day 54-55 , where the male rats were sexually mature , also cryptorchidism were seen as well as a dose - related decrease in penis length .

The results support that a change in AGD may predict permanent malformations of the external genitalia in rats later in life. Anogenital distance seems to be a reliable biomarker for later malformations and decreased penile length. Malformations of male offspring's genitalia could be scored early after birth (day 0 and day 6 in particular) and these early abnormalities turned out to be permanent and seemed to be deteriorating/worsening with age.

Results are seen below (unpublished)


External malformations PD1
120
100
80
60
40
20
0
% offspring
☆
☆☆
☆☆
Score 0
Score 1
Score 2
Score 3
1:Control
2: Procymidon 25
3:Procymidon 50
4: Procymidon 75
Score per dose group


External malformations PD6
120
100
80
50
40
20
0
% offspring
☆☆
☆☆
☆☆
Score 0
Score 1
Score 2
Score 3
1:Control
2: Procymidon 25
3:Procymidon 50
4: Procymidon 75
Score per dose group


13-23 malformation score external, PD 54-55
100
90
80
70
60
50
40
30
20
10
0
1: Control
2: Procymidone 25
3:Procymidone 50
4: Procymidone 75
Score 0
Score 1
Score 2
Score 3