Appendix 1: Study quality assessment and data extraction for *in vivo* studies on nipple/areola retention (NR)

Full-text publications on in vivo studies of effects on anogenital distance and nipple retention in mammals with exposure causing reduced testosterone (intratesticular or circulating) were analysed. Data from the publications were extracted into an Excel template, and the reliability of the studies was assessed using the SciRAP *in vivo* tool (*in vivo*: http://www.scirap.org). The specification of SciRAP evaluation criteria is presented in Table 1.

Studies were divided into different datasets for separate reliability evaluation if:
-Different chemicals were used
-Different exposure windows were used
-Different time points for observation of nipple retention were used

SciRAP evaluation criteria for the *in vivo* studies

The SciRAP tool (www.scirap.org) was used to assess the reliability of the included in vivo datasets. Only methodological quality (MQ) was assessed. In addition to the guidance for evaluation of each MQ criterion available in the SciRAP tool, specific considerations were developed and applied in this case and are listed in Table 1.
SciRAP MQ criteria 3, 11, 15, 16 and 17 were identified as especially critical for reliability in the current case and were selected as "key criteria" (Table 1).

Table 1. SciRAP evaluation criteria for the in vivo studies

Evaluation criteria, in vivo studies	**Comment on evaluation and handling in addition to online SciRAP guidance for each criterion**
Test compound and controls	
1. The test compound or mixture was unlikely to contain any impurities that may significantly have affected its toxicity.	Evaluated to be important for effects on nipple retention. Fulfilled: Purity is 98% or more. Partially fulfilled: 90-98% Not fulfilled: Purity below 90% - unless it is demonstrated that the impurities/contaminations do not affect the study outcome Not determined: Test compound purity is not reported. A comment is made to explain why.
2. An appropriate vehicle was used that is not expected to interfere with the absorption, distribution, metabolism, excretion, or toxicity of the test compound.	Evaluated to be potentially important, see TG 443 Fulfilled: Aqueous solution, solutions in oil Partially fulfilled: Acetone, DMSO, ethanol, tween-62, methylcellulose (they have potential intrinsic toxicity, but TG 443 still allows it).

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
3. A concurrent negative control group was included.	**Defined as a <u>key criterion</u>**
Animal model and housing conditions	
4. A reliable and sensitive animal model was used for investigating the test compound and selected endpoints.	Rats are most commonly used, but mice are also acceptable, for all strains.
5. Animals were individually identified.	For the parent animals, individual identification is judged to be potentially important. For the pups, individual identification is rarely seen and evaluated to be less important as long as the litters are identified as belonging to specific dams. This is usually the case, even if it is not reported. <u>Fulfilled:</u> If it is reported that parent animals were individually identified or housed separately. Whether it is reported or not that litters were identified as belonging to specific dams does not influence the judgement, since this is assumed to be the case unless the contrary is reported. <u>Partially fulfilled:</u> If there is no reporting of individual identification of dams. Whether it is reported or not that litters were identified as belonging to specific dams does not influence the judgement, since this is assumed to be the case unless the contrary is reported. A comment is made to clarify why the judgment was made. <u>Not fulfilled:</u> If the study description reveals that dams were not individually identified or if the study description reveals that litters were not identified as belonging to specific dams.
6. Housing conditions (temperature, relative humidity, light-dark cycle) were appropriate for the study type and animal model.	If the publication generally refers to the use of some (national/international) guidelines for the housing of animals, we assume that this is performed appropriately, and we judge it as “Fulfilled”. We make a comment that

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	details are not reported but that the publication refers to a specific guideline.
7. The number of animals per sex in each cage was appropriate for the study type and animal model.	Rats are social animals and should generally not be housed alone, except during the later part of gestation when they become more territorial. During this period and in the postnatal period until weaning, each dam/litter should, therefore, be separated and housed alone. The number of same-sex animals per cage is evaluated as potentially important since maternal stress may affect the sexual development of offspring, including nipple retention. In many rat studies, dams are housed in pairs until GD17 and alone thereafter. Fulfilled: Dams are housed in pairs until separation around GD17 or a few days before or after with some justification. Partially fulfilled: If the animals are housed alone or separation is introduced earlier in gestation without a justification. Not determined: If the number of animals per sex per cage is not reported. Make a comment to explain why.
8. The test system was unlikely to contain contaminants that could affect study results, such as organic pollutants, pesticide residues, heavy metals, and mycotoxins, as well as phytoestrogens.	Evaluated to be potentially important for effects on nipple retention if the test system contains ED substances. Regarding polycarbonate cages (PC), they may release small amounts of weak estrogenic substances and may influence results when testing estrogenic substances. However, for strong anti-androgens, possible exposure from PC cages is evaluated to be less important. Partially fulfilled: Cages are made of polycarbonate or other similar plastic material. Not determined: Cage type is not reported.
Dosing and administration of the test compound	
9. The allocation of animals to different treatments was randomized.	Random allocation into exposure groups is evaluated to be important, but it is also important for the proper conduct of a toxicity study that the body weight distributions between exposure groups

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	(at the beginning of the study), are similar. Therefore, "pseudo-randomization", a method where animals are not selected completely random but where the similarity in mean body weight between groups is obtained, is regarded as equally acceptable. Fulfilled: Complete randomization or pseudo-randomisation Not determined: Allocation of animals is not reported. Make a comment to explain.
10. The route of administration was appropriate and not likely to interfere with the study results.	Fulfilled: Oral (diet, drinking water, or gavage), dermal, and inhalation. Partially fulfilled: Subcutaneous administration (known to bypass liver metabolism). Other routes of administration are judged individually (as partially or not fulfilled), and a comment describing the exposure route is added.
11. The timing and duration of administration were appropriate for investigating the included endpoints.	**Defined as a key criterion.** The exposure period should include the male programming window, meaning gestation day 14-17 days post coitum in mice and gestation day 15-19 days post coitum in rats. Fulfilled: Exposure during all of the male programming window, i.e. GD 14-17 post coitum in mice and GD 15-19 post coitum in rats. Partially fulfilled: Exposure during some of the male programming window, i.e. GD 14-17 post coitum in mice and GD 15-19 post coitum in rats. A note is made to explain which period of exposure occurred. Not fulfilled: Exposure not in the male programming window.
12. The frequency of administration was appropriate for investigating the included endpoints.	One or a few exposures in the appropriate exposure period may be adequate for investigation but will complicate dose extrapolations/comparisons between studies. Therefore, daily dosing is preferable compared to other dosing

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	scenarios. Fulfilled: Daily dosing every day in exposure period chosen in study. Partially fulfilled: Not every day in exposure period chosen in study.
Data collection and analysis	
13. The allocation of animals to different tests and measurements was randomized.	Evaluated to be potentially important. Fulfilled: If it is reported that it was randomized or that every pup was measured. Not determined: If it is not reported that it was randomized or that every pup was measured. Make a comment to explain why.
14. Reliable and sensitive test methods were used for investigating the selected endpoints.	Nipple retention: It is very important that nipple retention is evaluated blinded to the exposure group. Additionally, NR should ideally be assessed by the same person throughout a study, and the person performing the analyses should be experienced, otherwise, the variation in data may increase substantially. Fulfilled: If it is reported that NR was investigated blinded and either by the same person OR by an experienced person. Partially fulfilled: If there is information indicating that it was investigated blindly but there is no information about whether it was investigated by the same person or by an experienced person. It is also judged as partially fulfilled if none of it was reported. A note is made to explain what was reported and what was not reported. Not fulfilled: It can be interpreted from the description or results that NR was not investigated blinded by the same person OR by an experienced person. Testosterone levels: There were no specific criteria for the measurements of testosterone, apart from that it should be measured ex vivo or in testicular homogenates.

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
15. Measurements were collected at suitable time points in order to generate sensitive, valid, and reliable data.	**Defined as <u>Key criterion</u>** Postnatal days 12-14 is the correct period for examination of nipple retention in male rat pups. Earlier than this, the areolae are not yet visible through the skin and later than this, the animals grow fur, and you need to shave them for proper examination. <u>Fulfilled:</u> NR is investigated between PND 12-14 or if investigation is initiated before PND12 and continued into PND12-14. <u>Partially fulfilled:</u> NR is investigated PND 15-22 and it is reported that the animals were shaved before assessment, or NR is investigated PND 10-11. <u>Not fulfilled:</u> NR is investigated before PND10 or NR is investigated PND15-22, and it is not reported that the animals were shaved. Investigation of NR after PND22 will not be included in the analyses since such results reflect a separate endpoint (they may inform on the severity of the exposure, but lack of permanent nipples should not be seen as a reason to dismiss observations of male areola identified on PND 12-14).
16. A sufficient number of animals per dose group were subjected to separate tests/data collection/measurements to generate reliable and valid results.	**Defined as a <u>key criterion.</u>** This criterion has two parts, as the group size needs to be sufficient both at the level of the number of litters per group and at the level of examined pups per litter. <u>Fulfilled:</u> Min. 8 litters/group AND there is no clear indication that only one male pup per litter was investigated. <u>Partially fulfilled:</u> 6-7 litters/group AND there is no clear indication that only one male pup per litter was investigated. The criterion is also partially fulfilled if there is toxicity in the highest dose group, leading to a lower number of litters or pups in that group (but the rest of the dose groups meet the criterion as fulfilled or partially fulfilled).

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	Not fulfilled: 1-5 litters/group OR clear indication that only one male pup per litter was investigated.
17. The statistical methods have been clearly described and do not seem inappropriate, unusual, or unfamiliar.	**Defined as a key criterion.** The litter should be regarded as the statistical unit. Fulfilled: If it is reported that the litter was used as the statistical unit, or the litter was taken into account in the statistics. Partially fulfilled: If it is not reported but can be interpreted from the results that the litter was used as the statistical unit. Not fulfilled: If it is not reported, and it can be interpreted from the results that the litter was not used as the statistical unit. It is also not fulfilled if no statistical tests were used.