Appendix to AOP Wiki text KER 3348

SciRAP evaluation criteria for 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 evaluating 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 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 have significantly 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 If test compound purity is not reported, this is judged as "not determined", and 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. (In TG 443: prioritised order: aqueous solution, solution in oil; vehicles with potential intrinsic toxicity should be avoided (e.g. acetone, DMSO) If acetone, DMSO, ethanol, tween-65 or methylcellulose is used, it is judged as "partially fulfilled" (TG 443 still allows it)
3. A concurrent negative control group was included.	Defined as a **key criterion**
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, all strains.
5. Animals were individually identified.	For the parent animals, individual identification is judged to be potentially important. Individual identification is rarely seen and evaluated as less important for the pups as long as the litters are identified as belonging to specific dams. This is usually the case, even if it is not reported.

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
	Fulfilled: 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 judgment, since this is assumed to be the case unless the contrary is reported. Partially fulfilled: 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 judgment, since this is assumed to be the case unless the contrary is reported. A comment is made to clarify why the judgment was made. Not fulfilled: 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 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. If the separation is introduced a few days before or after GD17 with some justification, the criterion is judged as "fulfilled".

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
	If the separation is introduced earlier in gestation without a justification, the criterion is judged as "partially fulfilled". If the number of animals per sex per cage is not reported, we leave it as "Not determined" and 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. If cages are made of polycarbonate, the criterion is judged as "partially fulfilled". We judge it as "Not determined" if it is not reported.
Dosing and administration of the test compound	
9. The allocation of animals to different treatments was randomised.	Random allocation into exposure groups is evaluated to be important. Still, it is also important for the proper conduct of a toxicity study that the body weight distributions between exposure groups (at the beginning of the study) are similar. Therefore, "pseudo-randomisation", a method where animals are not selected completely randomly but where the similarity in mean body weight between groups is obtained, is regarded as equally acceptable. If animal allocation into exposure groups is not reported, we report it as "Not determined" and make a comment to explain why.
10. The route of administration was appropriate and not likely to interfere with the study results.	The following routes of administration are judged to fulfil this criterion: Oral (diet, drinking water, or gavage), dermal, and inhalation. Other routes of administration are judged individually (as partially or not fulfilled), and a comment describing the exposure route is added. "Partially fulfilled" could be used for, e.g., subcutaneous administration, which is known to bypass primary liver metabolism.

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
11. The timing and duration of administration were appropriate for investigating the included endpoints.	**Defined as a <u>key criterion</u>.** 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 this will complicate dose extrapolations/comparisons between studies. Therefore, daily dosing is preferable compared to other dosing scenarios. Fulfilled: Daily dosing every day in the exposure period chosen in the study. Partially fulfilled: Not every day in the exposure period chosen in the study.
Data collection and analysis	
13. The allocation of animals to different tests and measurements was randomised.	Evaluated to be potentially important. If it is reported that it was randomised or that every pup was measured, we judge it as "fulfilled". If it is not reported that it was randomised or that every pup was measured, we judge it as "Not determined" and make a comment to explain why.
14. Reliable and sensitive test methods were used for investigating the selected endpoints.	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. Therefore, optimally, it is reported that experienced

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
	lab technicians (preferably one and the same person, blinded to exposure) investigated the nipple/areola assessment. Fulfilled: If it is reported that NR was investigated, blinded. It should also be reported that it is investigated blinded by the same person, OR by an experienced person. Partially fulfilled: If there is information indicating that it was investigated blinded, 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.
15. Measurements were collected at suitable time points in order to generate sensitive, valid, and reliable data.	Defined as **a Key criterion** 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, so you need to shave them for proper examination. Fulfilled: Therefore, this criterion is fulfilled if NR is investigated between PND 12-14. If the investigation is initiated before PND12 and continued into PND12-14, the criterion is also fulfilled. Partially fulfilled: If NR is investigated, PND 15-22, and it is reported that the animals were shaved before assessment. If NR is investigated PND10-11, it is judged as partially fulfilled. Not fulfilled: If NR is investigated before PND10. If NR was 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

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
	exposure, but lack of permanent nipples should not be seen as a reason to dismiss observations of male areolas 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 **key criterion.** 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. Fulfilled: Min. 8 litters/group AND there is no clear indication that only one male pup per litter was investigated. Partially fulfilled: -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). 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

Evaluation criteria, in vivo studies	Comment on evaluation and handling, in addition to online SciRAP guidance for each criterion
	unit. It is also not fulfilled if no statistical tests were used.

Inconsistencies in the empirical evidence for KER3348

In the evaluation of the WoE for DEHP, 13 datasets judged as reliable with restrictions (10) or without restrictions (3) were included. Nipple/areola retention was observed in 10 of 10 datasets judged as “Reliable without restriction”. In the three datasets judged as “Reliable with restriction” no statistically significant increases in nipple/areola retention were observed. In one of these, a 30-fold increase in nipple/areola retention was observed, but it did not reach statistical significance. This is not considered a conflicting result, even if the dataset is recorded as ‘no effect’. In the other two datasets judged as “Reliable with restrictions”, the lack of observed nipple/areola retention could be due to differences in exposure levels. The maximum doses used in these studies were 30 and 150 mg/kg/d, which are lower than the doses inducing effects in most other studies included; In 8 out of the 10 datasets judged as “Reliable without restrictions,” nipple/areola retention was observed with LOAELs of 300 mg/kg/d or above. The reliability of the two datasets judged as “Reliable with restrictions” was further hampered by key criterion 16 being only partially fulfilled, as only 5-7 litters were included in the exposed groups. This, along with the relatively low exposure doses investigated, could explain the conflicting results. Therefore, the conflicting results can be explained by dose and statistical power, and the overall level of confidence for prenatal DEHP exposure resulting in increased nipple/areola retention in male offspring is judged to be strong.

In the evaluation of the WoE for DBP, 7 datasets judged as reliable with (5) and without restrictions (2) were included. NR in male pups was observed in all 5 datasets judged as “Reliable without restriction”. In one of these, the effect was measured as number of pups with NR. In the other 4 datasets it was measured as number of areolas per. pup. A statistically significant increase in NR was observed in one of the two datasets judged as “Reliable with restriction”. In the other, a 20-fold increase in NR was observed but did not reach statistical significance. A relatively small group size was used in this study (N=6) and only one dose was tested. Even without statistical significance, a 20-fold increase is considered beyond normal biological variability and strongly suggests a treatment-related effect that deserves interpretation rather than dismissal. Therefore, the result is not interpreted as clearly “no effect” and no real conflict is observed between the two studies. Thus, the level of confidence for prenatal DBP exposure resulting in NR in male offspring is judged to be strong.