Appendix 1: Study quality assessment and data extraction for *in vivo* studies for KER-2828 'Decrease, AR activation leads to hypospadias'

Full-text publications on *in vivo* studies of effects on hypospadias in mammals with exposure to anti-androgenic substances were analyzed. 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 are presented in table 1 and the translation into reliability categories for each dataset was done using the principles laid in table 2.

Studies were divided into different datasets for separate reliability evaluation if different exposure scenarios led to assignment to different reliability categories

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 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, 14, 15, and 16 were identified as especially critical for reliability in the current case and were selected as "key criteria" (Table 1). For these criteria, the weight was increased in the SciRAP online tool.

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 significantly have affected its toxicity.	Evaluated to be important for effects on AGD 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

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	443 still allows it).
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 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. 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. If there are not reference to housing guidelines, but the conditions are evaluated as appropriate, it is marked as "Partially fulfilled".
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. Not fulfilled: If the animals are housed alone. 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 AGD 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. Not 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

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	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-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-randominzation 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.	The exposure period should include the male programming window, meaning gestation day 14-17 days post coitum in mice and gestation day 15.5-18.5 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: Other appropriate GD intervals 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 scenarios.
Data collection and analysis	

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
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.	**Defined as a key criterion.** Fulfilled: Hypospadias was assessed either by macroscopic examination or by proper histological analysis (I.e. coronal or sagittal sections in the middle of the GT). Partially fulfilled: The method of evaluation is not clearly described, but there are no indications that it is not reliable. Not fulfilled: The method of assessment seem flawed.
15. Measurements were collected at suitable time points in order to generate sensitive, valid, and reliable data.	**Defined as a key criterion.** Fulfilled: After ~PD24 in mice and rats (at the start of preputial separation). Partially fulfilled: ~PD1-PD24 (Here the penis is still developing). Not fulfilled: Assessment prior to PD1.
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.** Fulfilled: Min. 8 dams Partially fulfilled: 6-7 dams Not fulfilled: 1-5 dams. In studies, in which the highest dose groups had lower sample size than the other groups (e.g. due to toxicity), this criteria was judged according to the other dose groups.
17. The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar.	Statistics are rarely made on hypospadias incidences. Instead, this criteria is evaluated as follows: Fulfilled: The frequency of hypospadias in offspring is properly reported and if there are statistical analyses, they are appropriate. Partially fulfilled: No frequency of hypospadias is reported, but the descriptions of hypospadias imply a high frequency / canonical / hypospadias model (i.e. all animals have hypospadias).

Evaluation criteria, in vivo studies	Comment on evaluation and handling in addition to online SciRAP guidance for each criterion
	Not fulfilled: No reports on frequency of hypospadias and no justification for a "hypospadias model"

Table 2. Principles for translation of SciRAP scores to reliability categories.

Reliability Category	Principles
1. Reliable without restriction	SciRAP methodological quality score > 80 and all key criteria* are "Fulfilled," and there are no deficiencies in the non-key criteria that might affect study reliability.
2. Reliable with restriction	SciRAP methodological quality score > 65 and one or several of the key criteria are "Partially Fulfilled" or there are minor deficiencies in the non-key criteria that might affect study reliability.
3. Not reliable	SciRAP methodological quality score < 65 or one or several of the key criteria are "Not Fulfilled" or there are major deficiencies in the non-key criteria that affect reliability.
4. Not assignable	Two or more of the key criteria are "Not Determined"

*Key criteria are criteria judged as specifically critical for the reliability of the data in a certain case and are determined "a priori". The following five key criteria were used for *in vivo* studies: A concurrent negative control group was included, the timing and duration of administration were appropriate for investigating the included endpoints, reliable and sensitive test methods were used for investigating the selected endpoints, measurements were collected at suitable time points in order to generate sensitive, valid, and reliable data, a sufficient number of animals per dose group were subjected to separate tests/data collection/measurements to generate reliable and valid results.