Environmental Toxicology and Chemistry, 2025, **44(5),** 1435–1447
https://doi.org/10.1093/etojnl/vgaf067
Advance access publication: March 19, 2025
Original Article

Hazard/Risk Assessment

Amphibian studies to investigate the endocrine-disrupting properties of chemicals through the thyroid modality: a comparison of their statistical power

Simone Rizzuto[1,*], Franco Maria Neri[1], Valeria Ercolano[1], Alessio Ippolito[1], Alberto Linguadoca[1], Laura Villamar Bouza[1], and Maria Arena[1]

[1]European Food Safety Authority, Parma PR, Italy

*Corresponding author: Simone Rizzuto. Email: simone.rizzuto@efsa.europa.eu

S.R. and F.M.N. contributed equally to this work.

Abstract

Amphibians are the current model species for investigating the endocrine-disrupting (ED) properties through the thyroid modality in non-mammalian species. A recurrent question in the European Union (EU) regulatory endocrine assessment of pesticide active substances (2018/605) is whether the positive results from an in vivo screening test, that is, Amphibian Metamorphosis Assay (AMA), can be considered sufficient to conclude on the ED properties of a pesticide active substance or whether the larval amphibian growth and developmental assay (LAGDA) is a necessary step to further clarify the concerns identified in the AMA. Another one is the consideration of the extended AMA (EAMA). To further clarify some of the uncertainties around the use of the LAGDA and to help further consideration of the EAMA in a regulatory context, the statistical power of the three test designs was tested for all the parameters entailed to be measured in the respective study design (except for thyroid histopathology) by using data from real experimental studies. Our findings showed that the statistical power of the EAMA is in line with other Organisation for Economic Co-operation and Development standardized tests, that is, AMA, LAGDA. Our results also confirmed that the LAGDA is more powerful to detect effects on relevant parameters, that is, time to reach metamorphosis, compared to other in vivo tests. However, the difference in power was small, questioning its contribution to an overall weight of evidence already supporting the identification of a substance as an ED. These findings should be considered only in the context of hazard-based endocrine assessment of active substances (i.e., EU regulatory ED assessment of pesticide active substances, 2018/65), while they may not be fully applicable for a risk assessment-based approach.

Keywords: endocrine disruptors, amphibians, pesticide regulation

Introduction

Regulation 2018/605 (European Commission, 2018) lays down the scientific criteria for the identification of pesticide active substances having endocrine-disrupting (ED) properties. To support the implementation of those criteria, a guidance document was jointly developed by the European Chemical Agency (ECHA) and the European Food Safety Authority (EFSA) with the participation of the Joint Research Centre (ECHA/EFSA, 2018). In accordance with the Organisation for Economic Co-operation and Development (OECD) guidance document on standardised test guidelines for evaluating chemicals for endocrine disruption (OECD, 2018), the ECHA/EFSA guidance recommends using amphibian studies for investigating the ED properties in non-mammalian species through the Thyroid-(T) modality, because it is well established that amphibian metamorphosis is under the control of the hypothalamic–pituitary–thyroid axis. In case of uncertainties around some of the definitions used in this study, for example, "endocrine activity," "endocrine adversity," "sufficiently investigate," "T-modality," "T-mediated," "positive findings," "screening assay," and so forth, please refer to the ECHA/EFSA Guidance (2018) for clarification. Two standardized in vivo amphibian study designs are available to sufficiently investigate the endocrine activity and/or adversity of plant protection products through the T-modality. These studies are the amphibian metamorphosis assay (AMA, OECD test guideline 231, OECD, 2009; U.S. Environmental Protection Agency [USEPA], 2009) and the larval amphibian growth and developmental assay (LADGA, OECD test guideline 241, OECD, 2015; USEPA, 2015). A third assay, the Xenopus eleutheroembryonic thyroid assay (XETA, OECD test guideline 248, OECD, 2019) is also available. However, the XETA is only aimed at detecting thyroid active compounds based on induction of fluorescence compared to controls (Arena & van der Linden, 2021). Therefore, because the endpoints measured in the XETA are not comparable to those of the other available protocols and cannot be used to conclude on possible T-mediated adversity, it was not further considered in this study.

The OECD GD 150 also includes the OECD conceptual framework (CF) for testing and assessment of endocrine disrupters providing a grouping of the studies into five levels according to the

Received: November 12, 2024. **Revised:** February 25, 2025. **Accepted:** March 03, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of the Society of Environmental Toxicology and Chemistry.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

kind of information provided. According to this grouping, the AMA is a Level 3 CF mechanistic screening assay, which entails exposing *Xenopus laevis* tadpoles at stage Nieuwkoop & Faber (NF; 1994) stage 51 to a minimum of three different concentrations of a test chemical and control(s) for 21 days, collecting data for several responses (hereafter endpoints) at Days 7 and 21, that is growth parameters, developmental stage, and thyroid histology (only at test termination). The LADGA instead fits at Level 4 OECD CF, and endpoints are designed to characterize thyroid-mediated modes of action (MoAs), as well as estrogenic and androgenic. The test design of the LAGDA entails exposing newly spawned *X. laevis* embryos (NF stage 8–10) to a minimum of four different concentrations of the test chemical and control(s). When tadpoles reach metamorphosis (identified as NF Stage 62, hereafter "time to reach NF62"), a larval sub-sample is collected. After all animals have reached NF stage 66, that is, completion of metamorphosis (or after 70 days from the test initiation, whichever comes first), a cull is carried out at random to reduce the number of animals, and the remaining tadpoles continue exposure until 10 weeks after the median time to reach NF Stage 62 in the control(s), when additional measurements are made (see Table 1). The LADGA is the recommended test to sufficiently investigate the endocrine adversity for the T-modality, because it gives primary emphasis to potential population relevant effects. The ECHA/EFSA Guidance recommends following a tiered assessment strategy, starting with an AMA (or a XETA, under specific conditions reported in Annex A of the ECHA/EFSA Guidance, 2018) and in case of positive findings (in this context, significant findings indicating concerns of endocrine activity, according to the ECHA/EFSA Guidance, 2018), postulating a MoA analysis considering the endocrine adversity and/or adversity identified. At this stage, the need to generate further data may be considered (e.g., conducting a LAGDA), which could help analyzing the postulated MoA and to further assess/confirm the relevance of the adversity identified at the population level.

While the use of the AMA is well-established, the use of LAGDA in regulatory context is scarce and often criticized for several reasons. One is generally related to its validation, that is, limited number of chemicals, laboratories involved, and relevant MoAs for the thyroid. Another criticism is that while the exposure duration in the LAGDA is significantly longer compared to the AMA (i.e., 16 weeks vs. 21 days), the parameters considered to be diagnostic for the T-modality, that is, the time to reach NF 62, and thyroid histology, can only be collected at the interim larval sampling phase, similar to the AMA. This questions the added value of the LAGDA versus the AMA from a biological standpoint, particularly when a clear change compared to the control is observed in the AMA. However, while it is acknowledged that the sampling phase of the parameters is similar among the two test designs, the main difference is that the LAGDA is a more comprehensive test, specifically designed to collect concentration-response information on population-relevant adverse effects. In addition, the LAGDA has an unbalanced design (eight control replicates vs. four replicates in the treatments), intended to increase its statistical power. Therefore, the LAGDA is by-design inherently more powerful to detect significant effects compared to the AMA. However, to date, it is not clear to which level. This knowledge gap drove our first research question: what is the statistical power of the AMA and the LAGDA, and how do they compare? Answering this question would be instrumental to further clarify some of the uncertainties around the use of LAGDA to address concerns raised by positive findings in the AMA.

Recently, a revised AMA (extended or modified AMA—hereafter EAMA) testing methodology was proposed by Ortego et al. (2021). In such alternative design, the main experimental setup aligns with the criteria set in OECD test guideline 231, while the main change is in its duration. Specifically, the observation time where development is measured in the AMA is fixed to 7 and 21 days, when the test terminates (hereafter referred to as "fixed-time" design). Conversely, the EAMA has a "fixed-stage" design extending until tadpoles reach metamorphosis (i.e., stage NF 62). This latter design modification aligns with the methods used in the LAGDA. Such revision was considered to account for some limitations identified in the fixed-time design. In particular, a fixed-stage design may allow for a more straightforward data analysis and interpretation, avoiding confounded morphometric measurements and thyroid histopathology, because it could remove the potential influence played by pooling individuals at different developmental stages (Haselman et al., 2016, 2020; Olker et al., 2018; Ortego et al., 2014; U. S. Environmental Protection Agency, 2013). The authors of the EAMA identified greater statistical power to detect metamorphic delays compared to the fixed-

Table 1. Endpoints investigated in the amphibian metamorphis assay (AMA), extended (E)AMA, and larval amphibian growth and development assay (LAGDA).

	AMA (from NF stage 51, 21-day exposure)			Extended AMA (from NF stage 51 until stage NF stage 62)			LAGDA (from NF stage 8-10, 16 weeks exposure)		
Biological endpoints	**Daily**	**Day 7**	**Termination (Day 21)**	**Daily**	**Day 7**	**Termination NF stage 62**	**Daily**	**interim sampling (Larval), NF stage 62**	**Termination (juvenile), NF stage 66**
Mortality and abnormalities	×			×			×		
Wet weight		×	×		×	×		×	×
Hind limb length (HLL)		×	×		×	×			
Normalized hind limb length (nHLL)		×	×		×	×			
Snout to vent length (SVL)		×	×		×	×		×	
Developmental stage		×	×		×				
Time to reach NF 62						×		×	
Thyroid histology			×			×		×	
Organ histology (gonad, reproductive duct, gonads, kidney, liver)									×
Liver somatic index									×
Sex ratio									×

Note. NF = Nieuwkoop and Faber.

time design (Ortego et al., 2021). Nevertheless, some uncertainties remain because the statistical power of the EAMA was tested considering only delaying effects, and the comparison of developmental stage versus time to NF62 was carried out assuming a defined duration of developmental stages in days. This method may have some limitations, because the duration of developmental stages in days is variable and potentially influenced by many external factors. Therefore, the presence of such unanswered uncertainties drove our second research question: what is the statistical power of the EAMA and how does it compare with the AMA and LAGDA? Answering this question would be crucial to clarify the uncertainties around the EAMA, helping its further consideration for use in regulatory context, which, despite not being validated and not being recommended in the ECHA/EFSA Guidance (i.e., study design not yet available at the time of drafting), is already being submitted as part of dossiers for pesticide active substances' approval in the European Union (EU; Regulation European Commission [EC] No 1107/2009—European Commission, 2009).

To answer the two questions, the statistical power of the three test designs, that is, AMA, EAMA, and LAGDA, was tested by using data from real experimental studies. The analysis was conducted on all the parameters providing information on the T-mediated adversity. The thyroid histopathology was not considered in this study because, besides from the difficulties in the analysis owing to a complex data structure and lack of individual data in some cases, in the absence of significant effects in the other parameters, it provides information only on the endocrine activity. It is, however, acknowledged that thyroid histopathology is the most sensitive parameter in the case of thyroidal disruptors.

Materials and methods

Data collection and extraction

The data retrieval for the three test designs was carried out from study reports submitted to the EFSA in the context of pesticide active substances' approval (European Commission, 2009) for AMA (18 studies) and EAMA studies (9 studies), while for LAGDA (5 studies), data were sourced from the open literature, because no studies were yet submitted in the context of the ED assessment of pesticide active substances in the EU. For the AMA and EAMA studies, raw data were manually extracted from the available study reports, while for the LAGDA, data were provided by the corresponding authors of the published studies (Fort et al., 2017; Haselman et al., 2016, 2018). Data extraction was done on the endpoints generally collected in the three test designs, that is, total length, wet weight, snout-to-vent length (SVL), hind limb length (HLL), normalized hind limb length (nHLL), developmental stage or time to reach NF stage 62 (time to reach NF62; Table 1).

During data curation, for AMA designs, length data were discarded for animals exceeding NF stage 60, following the procedures recommended in OECD 231 (OECD, 2009). Comparison of the statistical power across the three test designs was not possible for the endpoint total length (only optionally measured in the AMA), because it was not recorded in the EAMA and LAGDA studies (data not shown). Comparison for the parameters HLL and nHLL was possible only between the AMA and the EAMA, because they are not recorded in the LAGDA (Table 1).

Statistical analysis

Power analysis for growth parameters

The continuous variables wet weight, SVL, HLL, and nHLL (see Table 1) were the focus of this analysis. As timepoint for comparison, Day 21 was used for the AMA, test termination for the EAMA, and interim sampling point (larval NF62) for the LAGDA. The assessment on Day 7 was initially included in the analysis but not finally considered for comparative analysis, given that the LAGDA studies did not include this intermediate timing. A comparative assessment of the variables HLL and nHLL was not possible for the three test designs because they are not measured in the LAGDA.

Prior to the analysis, departures from normality were assessed by visual inspection of linear model residuals, and only minor deviations from normality were found (see online supplementary material Figure S7). In parallel, a quantitative estimation of the deviation from homoscedasticity was performed by plotting the empirical cumulative distribution function of the p-values from the Levene's test (see online supplementary material Figure S6). The likelihood of significant departure from homoscedasticity was visually assessed as low. Altogether, the general assumption of normally distributed data and homogeneous variance across treatment levels was considered supported (see online supplementary material Text S5).

In line with the recommendations of the test guideline of the AMA (OECD, 2009), the no observed effect concentration (NOEC) was estimated for each response variable—time point combination using the Dunnett's test on individual observations ($\alpha = 0.05$). It is reminded here that while it is possible to extract a NOEC from the AMA and EAMA, it should not be used to set an endpoint to be used in risk assessment. For the LAGDA, it was noted that a mixed-modeling approach is recommended to explicitly consider within-replicate correlation (OECD, 2015). A power analysis using these two approaches would likely reveal measurable differences. However, replicating this analysis across the two models was considered out of scope, given the comparative nature of our assessment.

The minimum detectable effect size at the NOEC was measured using the proportional minimum detectable difference (pMDD; Brock et al., 2015; European Food Safety Authority [EFSA] Panel on Plant Protection Products and their Residues [PPR], 2013) and the proportional upper bound of the confidence interval (pCI; Mair et al., 2020). The Cohen's d (Cohen, 1988) was additionally calculated to provide insights on the estimated effect size at the NOEC. The pMDD and pCI were compared with a threshold set at 10%. Such value was selected because EC10 is often relied upon for assessing chronic effects (see and EFSA Panel on PPR, 2013) . Furthermore, it has been demonstrated that NOECs derived for the risk assessment of aquatic vertebrates (i.e., fish) tend to be, on average close to 10% effects (see Azimonti et al. 2015). Cohen's d values were compared to the standard effect size thresholds, that is, 0.2 = "Small effects"; 0.5 = "Medium effect" (Cohen, 1988). The influence of the test design on the MDD size in each study-response variable combination was estimated by using linear models with pMDD, pCI, or Cohen's d as response and test design as explanatory variable. Where relevant, a post-hoc comparison of Tukey's contrast was performed to quantify the differences among multiple design types. Residuals were checked for model validation purpose and for confirmation of the model assumptions of assuming normality and homoscedasticity. Finally, data were evaluated and graphically represented for the comparative analysis of test designs (Figure 1).

Power analysis for developmental stage (AMA)

For the analysis of developmental stage, the AMA recommends the multi-quantal Jonckheere-Terpstra (MQJT) test, which builds on the Jonckheere-Terpstra (JT) test. The JT test is a non-parametric test designed to detect a monotonic trend across

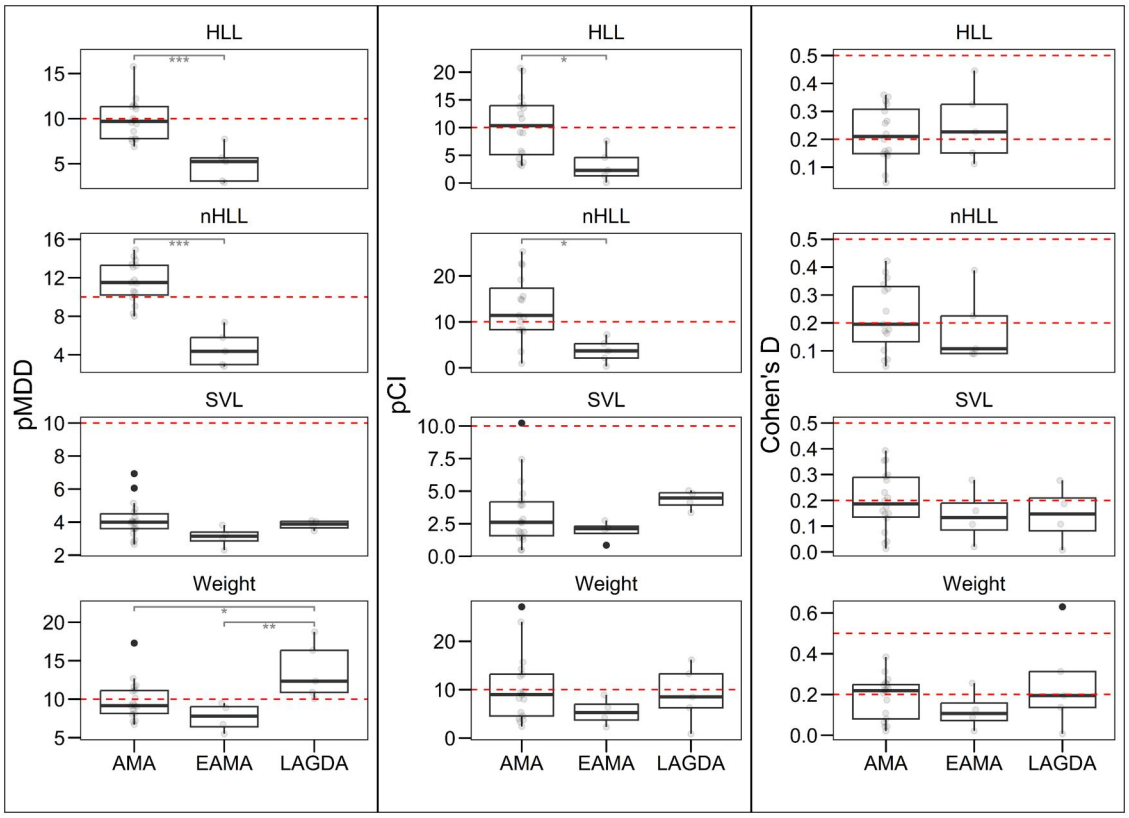


Figure 1. Proportional minimum detecatable difference (pMDD), proportonal upper bound of the confidence interval (pCI), and Cohen's *d* of the continuous variables investigated in the amphibian metamorphosis assay (AMA), extended (E)AMA, and larval amphibian growth and development assay (LAGDA). Comparative assessment of hind limb length (HLL) and normalized (n)HLL was not possible for the LAGDA because this parameter is not measured. Horizontal red dashed lines indicate the threshold of 10% for pMDD and pCI, while for Cohen's *d*, they highlight the standard thresholds for small (0.2) and medium (0.5) effect size. Black dots represent data points outside 1.5 times the interquartile range. The grey annotations indicate the statistically significant differences between test designs (*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$). SVL = snout to vent length.

groups (see Green et al. (2018) for details). It is a rank test, hence appropriate for ordinal data such as developmental stages; however, it does not account for replication. If replication is considered relevant, an option is to apply the JT test to a statistic (e.g., the median) extracted from each replicate, but this results in loss of information about within-replicate distribution (Green et al., 2018). A multi-quantal version, the MQJT test, was proposed for AMA studies to address these limitations. In the MQJT test (OECD, 2006), seven different quantiles (20%–80% at 10% intervals) are calculated for each replicate and treatment group of an AMA study. A standard JT test for trend across treatment groups is then applied independently for each quantile, resulting in a set of seven *p*-values. The median of the seven values is taken as the *p*-value of the MQJT, so that overall conclusions on trend are based on the majority of the quantiles. A statistically significant *p*-value of the MQJT test is considered as an indication of increasing/decreasing trend across treatment groups, that is, of a treatment effect (quantile-specific *p*-values may indicate trends in opposite directions, which requires more careful interpretation; this issue will not be addressed here). For the identification of the NOEC, the MQJT test is applied in a step-down fashion. It is first applied to the whole dataset, including the control and all treatment groups. If the outcome is statistically significant, the highest treatment group is dropped, and the MQJT test is applied to the reduced dataset. The procedure is iterated until the outcome is not significant: the highest at this point is the NOEC.

In the present analysis, the power of the MQJT test was estimated using control data from 17 AMA studies; an additional study did not include information on replicates and was dropped. The only feasible approach for this type of analysis is simulation-based for the following reasons. Estimating the power of a standard JT test is already considered a non-trivial problem. Closed-form expressions exist for the mean and variance of the JT statistic (Green et al., 2018), but an asymptotic normal approximation is discouraged for small sample sizes such as in the present case. For the MQJT test, this is further complicated by the fact that the outcome is based on seven JT tests, and the seven underlying datasets are correlated with each other (different quantiles of the same replicates). In addition, the test is applied in a step-down procedure for NOEC identification; hence, if the effect is on two or more treatment groups, the power is based on a sequence on MQJT tests, not just one. An extensive, simulation-based power analysis for the step-down MQJT test was presented in OECD (2006), Annex 5, and is still, to date, the most complete available analysis. As for the present study, the simulation in OECD (2006) was based on the generation of artificial datasets, with superimposed treatment effects of various kinds. In that case, however, the datasets were generated by sampling from parametric distributions (e.g., normal) varying the range of parameters. The approach followed here was different: the datasets were generated by resampling the control data of the available studies. The analysis was done using data for the negative control groups, as negative controls were available for all the studies. Solvent controls, which were included in a few studies, were not considered for the main analysis. The effect of using both control groups is

investigated in Text S3 (see online supplementary material Figure S4).

All the resampled datasets in this analysis consisted of a control group (treatment 0) and three test groups given increasing treatments (arbitrarily labeled 1, 2, and 3) of a hypothetical substance, with four replicates per group and 15 animals per replicate (whether or not there were missing animals in the original data). The procedure for each study consisted in repeating the following three steps.

Step 1: Generate a dataset with no treatment effect

Fifteen data points were randomly sampled (with repetition) from each replicate of the control group to create an artificial control dataset (four replicates with 15 data points each) where variability between replicates is preserved. The process was repeated four times, and the resulting data were labeled as treatment groups 0, 1, 2, and 3. The result was an artificial data set (240 data points) consisting entirely of resampled control data, that is, without treatment effect.

Step 2: Add a treatment effect

A treatment effect was generated by shifting the data of one or more test groups in the positive or negative direction (for accelerated and delayed development, respectively). Three main scenarios for the effect were considered, depending on the mean effect size (number of stages shifted) and the treatment groups affected (hence, on the no effect concentration, NEC): Scenario A. Mean effect size = 1 for treatment Group 3 (highest treatment group shifted by one stage): NEC = treatment 2. Scenario B. Mean effect size = 2 for treatment Group 3 (highest treatment group shifted by two stages): NEC = treatment 2. Scenario C. Mean effect size = 1 for treatment groups 2 and 3 (highest and second highest treatment groups shifted by one stage): NEC = treatment 1.

In addition to a given mean effect size, the full specification of an effect requires an assumption on inter-animal variability in the response, as the individuals in a group may not be all equally affected by the substance (Green et al., 2018). For example, a mean effect size = 1 (scenario A) in the negative direction corresponds in the simplest case to an experiment with uniform response (no variability) where all the animals in treatment Group 3 are equally delayed by one developmental stage. However, the same scenario may also describe a strongly bimodal response, where half of the animals in treatment Group 3 are unaffected and the other half are delayed by two developmental stages, resulting in a one-stage delay on average. The MQJT test may give different results for these two cases, and a power analysis should also investigate the possible impact of individual variability. For this purpose, in the present analysis, the effect was described as a random variable sampled from a distribution with a mean equal to the effect size and taking different values for each individual of the relevant treatment groups. Several discrete distributions, with different degrees of variability, were compared within each scenario.

Step 3: MQJT test

A two-sided MQJT test for trend ($\alpha = 0.05$) was applied to the resulting data set. The p-value of the test was calculated using exact permutation methods (Good, 2013). A sequential step-down procedure was used: in case of statistically significant trend, treatment Group 3 was dropped from the data and the MQJT repeated; the process was iterated until the outcome was not significant, and the highest treatment group at this point was considered the NOEC.

The output of steps 1–3 is a NOEC. The procedure in steps 1–3 was repeated 10,000 times for each study and combination of mean effect size (scenario), direction of effect (increasing/decreasing), and choice of distribution for the individual response. For each of those combinations, the power of the MQJT test was calculated as the proportion of simulations where the NOEC matched the NEC (number of simulations with NOEC = NEC divided by 10,000). Sensitivity analyses were carried out when relevant.

Power analysis for time to NF stage 62 (LAGDA and EAMA)

The purpose of the analysis was to compare the statistical power for the endpoint time to NF stage 62 between LAGDA and EAMA studies. The negative control data from nine EAMA studies and five LAGDA studies were used. All the control groups of EAMA studies consisted of 4 replicates and 15 individuals per replicate. Larval amphibian growth and development assay studies had a larger sample size and a more variable design: the replication of control groups was either 4 or 8, and the replicate size varied between 20 and 30. A few individuals who died before the end of the experiment (validity criteria identified in OECD 241 were still met) were not included in the analysis, resulting in a slightly smaller effective sample size for some of the studies and replicates. The main analysis was conducted for a two-sample comparison between the control and a hypothetical treatment; additional checks were done to extend the results to a comparison with multiple treatments. An additional check was also done for a few EAMA studies where two (negative and solvent) control groups were included: the two controls were compared and, if there were no significant differences, pooled for the power analysis, comparing the results with those obtained with negative controls only.

The analysis was done using two different approaches. The first approach was based on the Cox proportional hazard (PH) model, recommended by OECD test guideline 241 for the analysis of time to NF stage 62 in LAGDA studies. In the PH model, the effect is defined as a hazard ratio: a relative increase/decrease of the hazard (the instantaneous probability that an event occurs) between the control and the treatment. The key assumption of the PH model ("proportionality assumption") is that the hazard ratio is always constant during the experiment. The results of a power analysis can then be expressed in terms of a "minimum detectable hazard ratio" (MDHR, similar to the MDD): the smallest ratio that would be detected as significant with a given predefined power and confidence. In the present study, we considered the control data of each study and calculated the MDHR for a control-treatment comparison where the hypothetical treatment differs from the control by a given hazard ratio. The calculation was based on the sample size formula by Schoenfeld (1983). The significance level and power were set to $\alpha = 0.05$ and $1\text{-}\beta = 0.8$, respectively, for a two-tailed test. The power depends on two experiment-specific parameters (Schoenfeld, 1983): the allocation ratio, that is, the proportion of data control: treatment, and the overall number of events (individuals in the control and treatment groups that reach NF stage 62 before the end of the study). The MDHR was calculated using the allocation ratio specific to each study (1:1 for all EAMA studies; variable for LAGDA studies). The hypothetical number of events for the treatment was set equal to the number for the control, rescaled by the allocation ratio.

The second, alternative approach was based on the restricted mean survival time (RMST; Royston & Parmar, 2013). The

motivation for this choice was to address some of the disadvantages of the PH model: above all, that the proportionality assumption may be violated in practice, in which case the results of a power analysis would not be useful. In addition, even if the assumption holds, the outcome of the analysis (a hazard ratio) does not provide a readily interpretable summary of the effect. The RMST (Royston & Parmar, 2013) is defined as the average value of the time to event calculated up to a specific time point; it can be calculated with standard non-parametric methods of survival analysis (Eaton et al., 2020). In a two-sample comparison, the treatment effect can be expressed as a difference in RMST—the average shift in time (e.g., days) between the treatment and the control during the experiment, a straightforward measure of delayed/accelerated development. For the present study, we defined a MDD in RMST (MDD-RMST), the minimum shift in time that would be detected in a hypothetical comparison treatment–control. The MDD-RMST was estimated from control data ($\alpha = 0.05$ and $1\text{-}\beta = 0.8$ for a two-tailed test) with the following procedure. First, for each control dataset, we obtained a non-parametric estimate of the RMST (calculated between the first and last days of the experiment) using the Kaplan-Meier method; the variance of the estimate was calculated using the Greenwood plug-in estimator (Eaton et al., 2020). The allocation ratio (see above) was multiplied by the control variance to obtain a hypothetical RMST variance for the treatment. The latter calculation is justified if the variability is the same for the treatment and control; sensitivity to this assumption was tested using the original data for the treatment groups. The variance of the difference in RMST between treatment and control, obtained as the sum of the two individual variance estimates, was used to calculate the MDD-RMST based on the asymptotic expression for the test statistic (Eaton et al., 2020). All the replicates were pooled for the analysis.

For both the PH model and the RMST, the power analysis was done based on two-sample comparisons. In practice, however, the datasets comprise multiple treatment groups, and (in the most frequent approach) the analysis is done by conducting all possible two-sample control–treatment comparisons. A multiplicity correction is then usually applied to the resulting set of *p*-values (equivalently, to the significance levels) to avoid an inflated type I error (false positive) rate. Such correction results in higher *p*-values (or lower significance levels), so that the power of the test is lower than the value calculated for a two-sample comparison. We considered how the estimated two-sample power would change in case of multiple-treatment testing, assuming a Bonferroni–Holm multiplicity correction was used. With this correction, the initial significance level ($\alpha = 0.05$) of each comparison is rescaled: if n_g is the number of treatment groups, an effect on the highest treatment corresponds to a significance level $0.05/n_g$; an effect on the second highest treatment to a level $0.05/(n_g-1)$; and so on. The MDHR and MDD-RMST were re-calculated using rescaled significance levels and compared with the original, two-sample values.

Statistical software

All the analyses were conducted using R statistical software (Ver. 4.3.1 [June 16, 2023 ucrt]—"Beagle Scouts"; R Core Team, 2023) with the following packages: "car" (Fox & Weisberg, 2019), "clinfun" (Seshan & Whiting, 2023), "coxme" (Therneau, 2024), "CorrBin" (Szabo, 2024), "cowplot" (Wilke, 2020), "DescTools" (Signorell, 2024), "DHARMa" (Hartig, 2024), "dplyr" (Wickham et al., 2022), "effsize" (Torchiano, 2020), "forcast" (Hyndman & Khandakar, 2008; Hyndman et al., 2024), "ggbeeswarm" (Clarke et al., 2023), "ggfortify" (Horikoshi & Tang, 2016; Tang et al., 2016), "ggplot2" (Wickham, 2016), "ggsignif" (Ahlmann-Eltze & Patil, 2021), "ggsurvfit" (Sjoberg et al., 2024), "multcomp" (Hothorn et al., 2008), "patchwork" (Pedersen, 2024), "pwr" (Champely, 2020), "RcppAlgos" (Wood, 2024), "readxl" (Wickham & Bryan, 2022), "RColorBrewer" (Neuwirth, 2022), "Rfast" (Papadakis et al., 2023), "RMSTdesign" (Eaton, 2025), "survival", "survminer" (Kassambara et al., 2024), "survMisc" (Dardis, 2022), "survRM2" (Uno et al., 2022), and "tidyr" (Wickham & Girlich, 2022).

Results and discussions

Power analysis for growth parameters—AMA, EAMA, and LAGDA

The pMDD and pCI values for the majority of the continuous variables were close to the threshold value of 10% for all three test designs (Figure 1), indicating that AMA, EAMA, and LAGDA can detect significant effects between controls and the treatments with a minimum difference in their means of around 10% (Brock et al., 2015; EFSA Panel on PPR, 2013; Mair et al., 2020). More specifically, pMDD informs on the effect size, which would have been marked as significant (with $\alpha = 0.05$) had the variability remained the same (i.e., the minimal difference between the means of the control and the treatment for the NOEC to become a lowest observed effect concentration). The pMDD is not affected by the actual effect size observed in the specific experiment, but just by the sample size and the sample variability. On the other hand, pCI represents the maximum level ($\alpha = 0.05$) of the "true effect" at the NOEC—as such, pCI depends both on the effect size and sample size/variability (Mair et al. 2020). In the AMA, the median pMDD and the pCI values for the nHLL were greater than the threshold value (of ca. 2%). The pMDD and the pCI for the HLL and the nHLL were significantly higher in the AMA compared to the EAMA ($p < 0.001$ and $p < 0.05$, respectively), while no significant differences were observed for the SVL between AMA and EAMA. This pattern of results indicates that the EAMA can capture smaller minimum differences for some parameters compared to the AMA, likely due to the different study design, that is, fixed-stage versus fixed-time. The analysis conducted on the key drivers of pMDD/pCI (see online supplementary material Text S6 and Figure S8) may bring further insights on the observed differences between EAMA and AMA. For instance, online supplementary material Figure S8 shows that in comparison to the AMA, the mean of the control group in the EAMA is consistently (i) higher for the HLL and nHLL, (ii) lower for the SVL and wet weight, while the pooled SD and the coefficient of variation (CV) are consistently lower in the EAMA for all parameters. This different pattern reflects both the biological changes observed in amphibians at metamorphosis and the potential advantages of the fixed-stage study design. In the EAMA, tadpoles are in a more advanced stage of development (NF Stage 62) compared to the AMA; therefore, some morphometric parameters, such as the HLL, may be bigger. At the same time, still due to the morphological changes occurring with metamorphosis (e.g., re-adsorption of the tail) in the EAMA, other morphometric parameters, for example, SVL and wet weight, may be lower. No correlation between HLL and SVL in late-stage (NF > 60) datasets was also observed by other authors (Pawlowski et al., 2019). Lower SDs and CVs values in the EAMA compared to the AMA reflect the lower variability of the parameters likely due to the synchronization of the tadpoles at stage NF62 compared to the fixed-time design when exposure is stopped after 21 days, supporting the hypothesis of other authors (e.g., Ortego et al.,

2021). Although HLL is not investigated in the LAGDA, similar results for this test design are reported for SVL (see online supplementary material Figure S8). Nonetheless, before drawing general trends for the LAGDA, this finding should ideally be confirmed by a larger dataset than the one available in the current study.

In the LAGDA, the pMDD for the wet weight was >10%, and also significantly higher compared to the values observed in both the AMA and the EAMA ($p < 0.05$ and $p < 0.001$, respectively). The analysis in online supplementary material Text S6 (see online supplementary material Figure S8) shows that the lower power for wet weight is determined by the sample size, as the variability (CV) of the parameter for the LAGDA is similar to the AMA and the EAMA. The sample size for continuous growth parameters is smaller for the LAGDA. In the AMA and EAMA, those parameters are measured on up to 15 animals per replicate (those remaining at termination); in the LAGDA, they are measured on five randomly sampled animals per replicate. This difference was not observed in the pCI results. An explanation on the different behavior of pMDD vs. pCI is given in the online supplementary material Text S6. The results from the Cohen's *d* indicate that the NOECs for AMA, EAMA, and LAGDA are generally associated with effect sizes ranging from small (≤0.2) to medium (≤0.5; Cohen, 1988) for all the investigated growth parameters (Figure 1). No significant differences were detected between the three test designs. Cohen's *d* provides a measure of the effect size relative to the variability of the underlying population, and it is frequently used to compare effect sizes between different studies. Cohen's *d* is also strictly related to the *p*-value of the test: as explained in online supplementary material Text S6, this is the reason why the values calculated in the power analysis fall approximately in the range 0–0.5. Within this range, there are no substantial differences between study designs, which can be considered as an indication that the groups of studies were approximately comparable in terms of effect size.

Power analysis for developmental stage (AMA)—impact of variance

The results of the power analysis based on resampled datasets are organized as follows. First, the outcome for scenario A (see *Materials and Methods* section) is discussed in detail to illustrate how the power changes across studies and for different degrees of individual variability in the response. This is followed by an overview of the results for all three scenarios (A, B, and C). Further results are provided in the online supplementary material: an analysis of bootstrapped quantiles of the original datasets to identify the possible drivers of the patterns identified for scenario A (see online supplementary material Text S1); a discussion of two sensitivity analyses (see online supplementary material Text S2); an investigation of statistical power when solvent control groups are included (see online supplementary material Text S3); and more details on the distribution of NOECs (see online supplementary material Text S4).

Figure 2 shows the results for scenario A, for a mean effect size of one stage (decreasing/increasing, different symbols in Figure 2) and for three different degrees of individual variability in the response (different colors in Figure 2) measured by the variance of the underlying discrete distribution. In the case with no variability (variance = 0, blue triangles), the final stage of each animal in treatment Group 3 is uniformly shifted by 1. Non-zero variability is represented by two bimodal distributions. The distribution with variance = 1 (orange triangles) takes values 0 and 2 with probability 1/2 each: a randomly picked half of the animals in treatment Group 3 is shifted by two stages and the other half is unchanged. The distribution with variance = 2 (yellow triangles) takes values 0 and 3 with probabilities 2/3 and 1/3, respectively (hence, 1/3 of the animals on average is shifted by three stages).

Two main results emerge from Figure 2. The first is that, even with uniform response (variance = 0, blue triangles), statistical power varies strongly across studies. For instance, in this case, the power ranges from 3%–4% (study ID 15; the values provided are for the two directions of the effect) to 84%–87% (study ID 17), with only four studies showing values >60%, only one reaching >80%, while the majority were below 40% (10/17). An analysis of the quantiles of the resampled control datasets (bootstrapped quantiles) was carried out to gain further insight into the origin of variability between studies (see online supplementary material Text S1 and Figure S1).

The second result is a clear impact of variability on statistical power, which can be seen by comparing increasing values of the variance (colors in Figure 2) for each study. For studies that have at least 40% power in the case of uniform response (variance = 0), there is a consistent decrease in power when the variance of the effect increases. The decrease tends to be stronger the higher the initial power, thus smoothening out the variability between studies. For studies with very small power (<40%), a consistent pattern could not be observed, likely due to the starting baseline being already low; however, this can be considered of less relevance. It is acknowledged that the two-value distributions

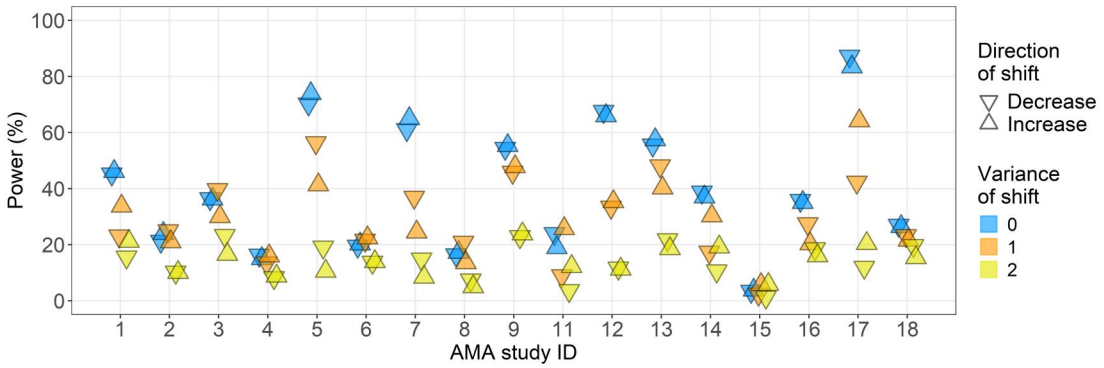


Figure 2. Results of the power analysis for scenario A, with a mean effect of one-stage for treatment Group 3 (highest treatment). The power is defined as the proportion of simulations that estimated a no observed effect concentration (NOEC) = 2 (matching the no effect concentration [NEC]). The study with ID 10 was omitted from the analysis because there was no information on the developmental stage of replicates. AMA = amphibian metamorphosis assay.

chosen for Figure 2 represent an extreme case. Several other distributions with the same values of the variance were tested and, although the emerging pattern was the same, the impact on statistical power was, in general, smaller (data not shown).

The observed variability in statistical power shown in Figure 2 was not completely unexpected, considering that the AMA is a screening study, designed to give a more "qualitative" rather than "quantitative" response, and therefore it is not recommended for the setting of a NOEC suitable for use in environmental risk assessment. Fixed-time design could have an indirect role in the observed different power across AMAs. Although the testing conditions (e.g., temperature, pH) in the AMA are standard, the developmental stage reached in 21 days by the tadpoles may vary considerably, ranging from NF57 to NF61 (data not shown). This is because other factors may impact the development of the tadpoles, such as, for example, animal source, rate and type of food, and iodide content. Potentially, such consideration could be made also for the EAMA, because a variation in some of the previously mentioned parameters could influence the number of days needed to reach NF62, although a maximum time is allowed (i.e., 45 days). Other "controllable" factors could be involved too, nevertheless the aim of this study is not to understand which factors may increase the statistical power of the AMA, and further research could be carried out into this direction. Results from the two extreme cases (ID15 and ID17) in Figure 2 (and online supplementary material Figure S1) showed how between- and within-replicate variability of the distribution of stages in controls may drive the results, further reinforcing the importance of ensuring adequate control, performance thereby reducing variability. Therefore, laboratories should take into account the influence of such parameters for more harmonized developmental data. Furthermore, presenting the power of the test to detect an effect on biometric parameters, for example, HLL, nHLL, and SVL, could be routinely implemented to increase the confidence of regulators.

Power analysis for developmental stage (AMA)—impact of effect size and NOEC setting

The results for scenarios A, B, and C are summarized in Figure 3. For ease of reporting, the studies were classified into three groups based on statistical power: best-performing studies with power ≥80%; an intermediate group with power between 50% and 80%; and the worst-performing studies with power ≤50%. The bar plots in Figure 3 show the number of studies in each category for an effect in the decreasing direction; the results for the increasing direction can be considered comparable (data not shown; however, data and scripts are available upon request). The distributions for individual response for scenario C (variance = 1, 2) are the same as for scenario A. For scenario B, bimodal distributions were also used: with variance = 4 (values 0 and 4 with probability 1/2 for both) and variance = 8 (values 0 and 6 with probability 2/3 and 1/3, respectively). In general, however, an impact of increasing variability can be observed for all scenarios, similar to what discussed for scenario A. Two comparisons between scenarios can be done: those with the same NEC and different effect size (A and B) and those with the same effect size and different NEC (A and C). In the first case, scenario B has twice the effect size as scenario A and, as a consequence, a much higher chance of detecting a significant trend, at least for small between-individual variability. As variability is increased, however, the differences between the two scenarios, for example, between counts of "useful" studies in the highest range, tend to disappear. From one side, these findings could be interpreted as a limitation of the test, considering that a significant two stages shift (either increase or decrease) is rarely observed with pesticide active substances, because they are not specifically designed to directly affect the thyroid, unlike other more powerful (anti)thyroidal pharmaceuticals, for example, methimazole, L-thyroxine (Coady et al., 2010; OECD, 2007). At the same time, the observed strong dependence of the statistical power on the effects size could highlight the fact that the AMA is designed to serve as a

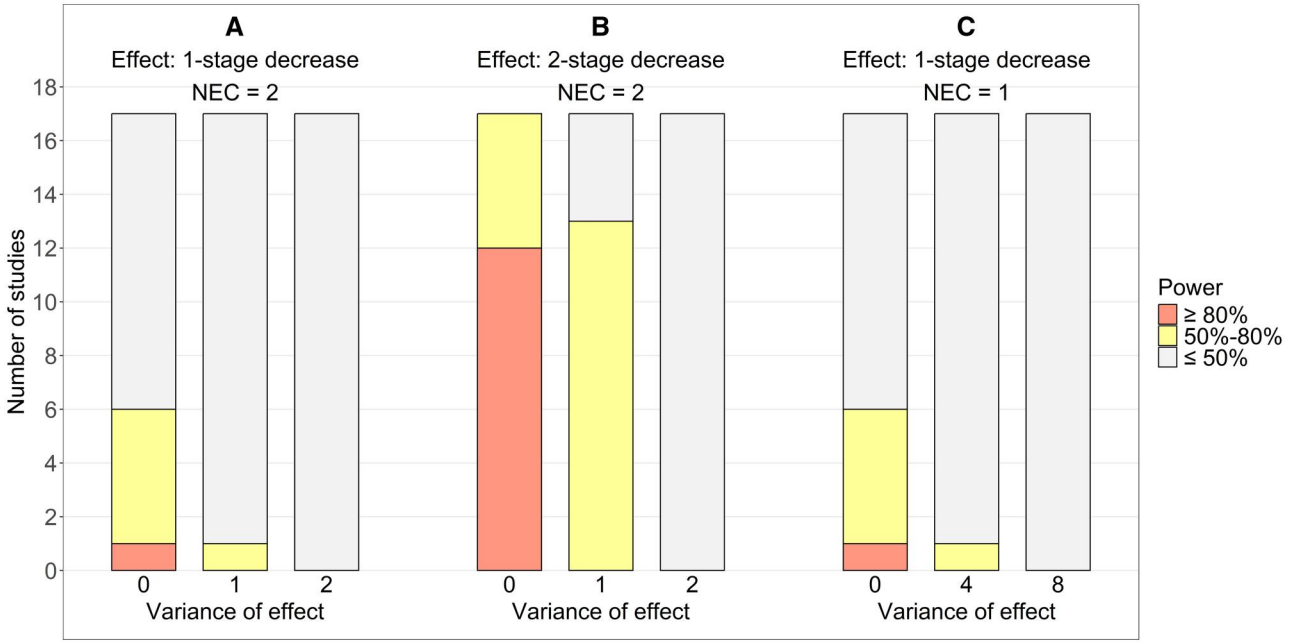


Figure 3. Overview of the results of the power analysis for the three main scenarios A, B, and C, based on a total of 17 amphibian metamorphis assay (AMA) studies. The studies are classified into different ranges (colors) of statistical power. A no effect concentration (NEC) = 2 corresponds to an effect on the highest treatment group (treatment = 3); a NEC = 1 to an effect on the highest and second highest treatment groups. The results are for the decreasing direction of the shift; those for the increasing direction are only slightly different (data not shown here but reproducible as included in the available R script).

screening test, and therefore, potential smaller effects such as one stage shift may already represent a concern that should be further clarified with higher tier experimental designs (i.e., LAGDA), especially if such effects are not observed in isolation, that is, accompanied by changes in thyroid histopathology.

The second comparison is between scenario A and scenario C: in the latter, the power is based on the detection of a NEC = 1, that is, the fraction of simulations that detected a significant trend both for the highest *and* the second highest treatment groups (see online supplementary material Text S4 and Figure S5 for the full distribution of NECs). In this case, the results are very similar for all values of the variance. The counts for the three main categories (bar plots) are identical for A and C, and inspection of the detailed results shows that the power for scenario C is only slightly lower. The comparison across scenarios shows that the power of the test is mainly determined by the number of stages, while the number of treatment groups affected has only a negligible impact. This outcome was already observed in the power analysis in OECD (2006), Annex 5, and further reinforces the fact that the AMA is not recommended for the setting of a NOEC suitable for use in environmental risk assessment (as also reported earlier). All the results presented in this section were based on data for negative control groups, as also reported in the *Materials and methods* section. Some studies also included data for solvent controls: an additional analysis was carried out for those studies, to investigate how the results would change by using two control groups. The analysis showed that by including the solvent control, there would be in general no gain, and often a loss of statistical power (see online supplementary material Text S3 and Figure S4). Two sensitivity analyses were also carried out, showing that additional noise had a clear impact on the studies with the highest power (see online supplementary material Text S2 and Figures S2 and S3).

Power analysis for time to NF stage 62 (LAGDA and EAMA)

The results of the power analysis for time to NF stage 62 are shown in Figure 4. The MDHR (Figure 4, left) was estimated from control data based on the Cox PH model. The MDHR is the minimum hazard ratio that would be detected as significant in case of an *increasing* effect on the hazard. For a *decreasing* effect, the equivalent measure would be a *maximum* detectable hazard ratio: this is the inverse of the MDHR, however, and it will not be discussed as the results are completely equivalent. The values of the MDHR are clustered at approximately 1.7 for EAMA studies and at approximately 1.4 for LAGDA studies, with very little variability within study type (EAMA/LAGDA). The differences between the two study types are determined by the sample size. As explained in the *Materials and* methods section, the power of the two-sample PH test increases (the MDHR decreases) with the number of events (individuals reaching NF stage 62 during the experiment): This is approximately proportional (almost equal) to the sample size, which is always larger for LAGDA studies.

The MDD-RMST (Figure 4, right) is the minimum shift in time (days) that would be detected in an analysis based on the RMST. The results are more variable than for the PH method. The range of values is 0.8–2.7 days for EAMA studies and 0.8–1.8 days for LAGDA studies, with considerable overlap. We remark that the

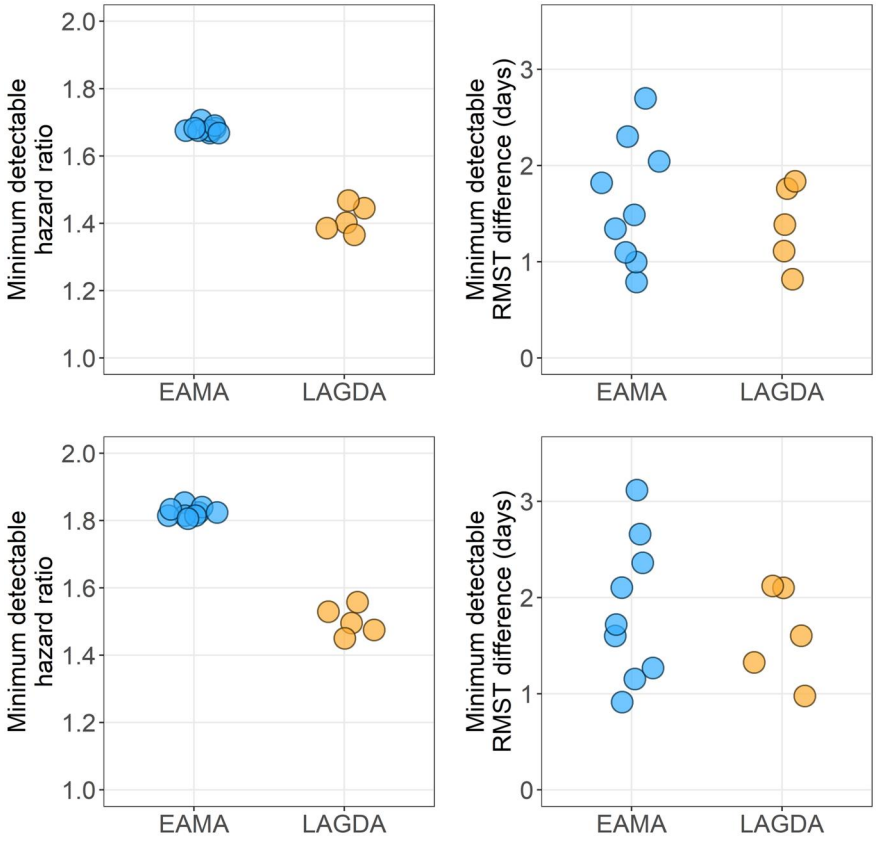


Figure 4. Results of the power analysis on the parameter time to reach Nieuwkoop and Faber (NF) stage 62 for nine extended amphibian metamorphis assay (EAMA) studies and five larval amphibian metamorphis assay (LAGDA) studies, based on two different approaches: the Cox proprtional hazard (PH) model (left) and the restricted mean survival time (RMST; right). The upper panels report two-sample comparison between the control and a single treatment, the ones below report multiple comparison between the control and all treatments groups followed by a Bonferroni–Holm multiplicity correction.

statistical power (expressed as MDD-RMST) depends on the variance of the RMST, which is calculated non-parametrically from the time-to-event curve (see *Materials and methods* section). Hence, in contrast with the MDHT, it is not possible to link the MDD-RMST to specific parameters of the experimental design. A strong assumption underlying the results in Figure 4 is that the variability calculated from the time-to-event curve of the control is the same for the treatment. Sensitivity to this assumption was tested using as a worst-case scenario a 1:3 control: treatment ratio for the variances (instead of 1:1). This was the maximum ratio estimated across all studies and treatment groups; note that, in general, the ratio varied below and above 1:1. Using a 1:3 ratio for all the studies, the MDD-RMST values in Figure 4 would increase approximately by a factor of 1.4.

The results in the upper panel of Figure 4 are based on a two-sample comparison between the control and a single treatment. An additional power analysis was done to extend the results to the case of a multiple comparison between the control and all treatments groups followed by a Bonferroni–Holm multiplicity correction (lower panels of Figure 4). In the additional analysis, both the MDHR and the MDD-RMST were re-calculated using rescaled significance level (starting from the initial $\alpha = 0.05$) for the detection of a hypothetical effect on a specific treatment. For an effect on the highest treatment only, the MDHR increased on average to 1.8 for EAMA studies and 1.5 for LAGDA studies; the MDD-RMST increased by a factor 1.2 for all studies. For effects on lower treatments, the changes were consistently smaller, with MDHR and MDD-RMST approaching the two-sample values as the number of treatments affected increased.

The results in Figure 4 were found using negative control data. A solvent control was included for four EAMA studies. For three of them, no significant differences were found between the negative and solvent control using either the PH model or the RMST, which justified pooling together the two control groups in the analysis. The results of the analysis with pooled controls showed a clear improvement in statistical power, with decrease in both the MDHR and MDD-RMST. For example, compared to the results in the upper panel of Figure 4, the MDHR with pooled controls changed by a factor 0.93 and the MDD-RMST changed by a factor approximately 0.6 for all three studies.

Based on these findings, both the test designs showed similar statistical power to detect significant effects on the parameter time to reach NF62 with an MDD-RMST range of 1–3 days, with the LAGDA showing slightly better performance compared to the EAMA, based on the results of the MDHR. These results show that the LAGDA is inherently more powerful than the EAMA simply based on the differences in sample size (640 vs. 400 tadpoles, respectively). For instance, the LAGDA is a more comprehensive test designed to collect concentration-response information on population relevant adverse effects, with a minimum of 640 tested tadpoles (at least 4 treatments, 4 replicates per treatment, 8 for the control(s)), while the EAMA is a modified version of a screening test (AMA) where a minimum of 3 concentrations are tested over 4 replicates with 20 tadpoles each (total of 400 animals). At the same time, the results from the MDD-RMST indicate negligible differences between the two tests in both two-sample and multiple comparison with the majority of LAGDA and EAMA showing similar results and only 2–3 EAMA deviating from such pattern.

AMA, EAMA, and LAGDA: can we compare developmental stage vs. time to NF62 parameters?

The results from the three test designs are expressed on very different times, with the AMA showing results in developmental stages (ordinal variable), while the EAMA and LAGDA as MDD in days. To make a direct comparison between AMA versus EAMA/LAGDA, stages would need to be mapped in days, as also proposed previously by Ortego et al. (2021). However, if such approach is to be taken, it would build on strong assumptions, without excluding the occurrence of several uncertainties, because the duration of stages in days is variable and may be influenced by many external factors. Therefore, considering the potential weight of such uncertainties on the outcome of the analysis, we preferred instead to consider a more minimalistic approach and avoid making a direct comparison by weighing the available data for an indirect evaluation. For instance, by taking into account the minimum "measurable" effects such as the shift in one developmental stage in the AMA, and of one day in both the EAMA and the LAGDA, all three test designs show similar pattern of distribution of the studies across the different levels of the statistical power. The results from the MDHR suggest that LAGDA may have a slight advantage over the EAMA (and indirectly over the AMA), confirming that such test may be better suited to detect adverse effects relevant at the level of population, which makes it also relevant to set a NOEC suitable to be used in environmental risk assessment. Nevertheless, the results from the MDD-RMST also showed negligible differences within the three test designs, which was unexpected considering the different levels of inherent statistical power between the LAGDA versus the other in vivo tests.

Conclusions

To further clarify some of the uncertainties (i) around the use of the LAGDA to address concerns raised by positive findings in the AMA and (ii) around the EAMA, helping its further consideration in regulatory context, the main aim of this work was to test the statistical power of the three in vivo test designs for all the investigated parameters, with the exception of thyroid histopathology, to answer the following research questions: What is the statistical power of the AMA and the LAGDA, and how do they compare? What is the statistical power of the EAMA, and how does it compare with the AMA and LAGDA? Based on the results obtained in this study, the following responses are presented.

In response to the first question, the statistical power of the AMA can be defined as sufficient to detect significant effects on growth parameters, performing even better than the LAGDA for the wet weight, although an exceedance from the set thresholds was observed for the pMDD and the pCI. The statistical power to detect effects (both advancement and delay) on developmental stage was not constant, was significantly affected by an increase in variance and by the effect size (i.e., number of stages affected), while the number of treatment groups affected has only a negligible impact. A direct comparison against the time to reach NF62 of the LAGDA was not made; however, based on a more minimalistic, indirect evaluation, the LAGDA showed a slight edge over the AMA.

In response to the second question, the statistical power of the EAMA could be considered in line with one of the OECD standardized tests, such as the AMA or the LAGDA. For some of the growth parameters, the statistical power of the fixed-stage design was higher compared to the AMA (HLL and nHLL) and the LAGDA (wet weight). The EAMA slightly underperformed compared to the LAGDA for the time to reach NF62 (as observed for the AMA vs. LAGDA). For this variable, based on the same minimalistic approach mentioned above, the power of the EAMA tests could be considered comparable to that of the AMA.

Overall, the findings from the AMA are consistent with the power analysis in OECD (2006), Annex 5, and further reinforce that this test is a screening study, designed to give a more "qualitative" rather than "quantitative" response, and therefore is not recommended for the setting of a NOEC suitable for use in environmental risk assessment. Our results also highlight the fact that a significant one-stage shift may already represent a concern that should be addressed with higher-tier experimental designs, unless the weight of evidence is already suggesting that the pattern may not be related to a disruption of the hypothalamic–pituitary–thyroid axis. In addition, based on our results, harmonization in the "controllable" factors (e.g., animal source, rate and type of food, iodide content) is recommended to help lowering intra- and inter-replicate variability, ensuring adequate control performances and likely allowing higher statistical power. The growth parameters' results from the EAMA suggest that the fixed-stage design may be more powerful compared to the fixed-time, likely due to the tadpoles being synchronized at test termination, compared to when exposure is stopped after 21 days. This conclusion aligns with the hypothesis of other authors (Haselman et al. 2016, 2020; Olker et al. 2018; Ortego et al. 2014; U. S. Environmental Protection Agency, 2013). This outcome, together with the results of the analysis on the time to NF62 parameter, can be considered reassuring for regulators and represents a step forward in terms of regulatory acceptance of this study design in the ED hazard assessment for non-target organisms. Nonetheless, to have a clearer picture, such results should be considered alongside the thyroid histopathology. Therefore, further research is required.

The LAGDA showed lower statistical power for the wet weight compared to both AMA and the EAMA, and at the same time had the highest power to detect effects on time to reach NF62 compared to the EAMA. The first outcome was explained with the lower sample size used in the LAGDA to investigate effects on SVL and wet weight (measured in only five of the tadpoles sampled at NF stage 62 per treatment group) compared to the EAMA (where measurements were carried out on 15 tadpoles). The second finding was expected because, when using the full number of tadpoles available per treatment group, the LAGDA is significantly more animal intensive with a higher number of control replicates compared to treatments and more treatments. This makes it better suited to collect concentration-response information on population relevant adverse effects and is relevant for environmental risk assessment purposes. At the same time, the magnitude of the observed difference compared to the other in vivo tests was unexpectedly small, questioning its contribution to an overall weight of evidence already supporting the identification of a substance as an ED, as also reported by other authors (Dang, 2022). It is acknowledged that these findings should be considered only in the context of hazard-based endocrine assessment of active substances (i.e., EU regulatory ED assessment of pesticide active substances, 2018/65), while they may not be fully applicable for risk assessment-based approach.

Supplementary material

Supplementary material is available online at *Environmental Toxicology and Chemistry*.

Data availability

Dataset and R scripts are available on Zenodo (10.5281/zenodo.14923622).

Author contributions

Simone Rizzuto (Data curation, Investigation, Project administration, Visualization, Writing—original draft, Writing—review & editing), Franco Maria Neri (Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft), Valeria Ercolano (Formal analysis, Methodology, Software), Alessio Ippolito (Formal analysis, Methodology, Software, Writing—review & editing), Alberto Linguadoca (Data curation, Formal analysis, Methodology, Software, Writing—review & editing), Laura Villamar Bouza (Writing—review & editing), and Maria Arena (Conceptualization, Supervision, Writing—review & editing)

Funding

No specific funding has been obtained for this publication.

Conflicts of interest

None declared.

Disclaimer

All the authors contributing to the present article are employed with the European Food Safety Authority (EFSA). However, the present study is published under the sole responsibility of the authors and may not be considered as an EFSA scientific output. The positions and opinions (if any) presented in this article are those of the authors alone and are not intended to/do not necessarily represent the views/any official position of EFSA. To know about the views or scientific outputs of EFSA, please consult its website under http://www.efsa.europa.eu.

Acknowledgments

The authors would like to express gratitude to the S.J. Degitz, T. Iguchi, D.J. Fort, and all co-authors for sharing the raw data of the LAGDA studies, and to A. Kienzler for the support provided in the data extraction. The authors also wish to thank the anonymous reviewers for their work.

References

Ahlmann-Eltze, C., & Patil, I. (2021). *ggsignif: R Package for displaying significance brackets for 'ggplot2'*. PsyArxiv. https://doi.org/doi : 10.31234/osf.io/7awm6

Arena, M., & van der Linden, S. (2021). Annex A—Use of the XETA in the assessment strategy of the ECHA/EFSA Guidance. Annex to Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA Journal*, *16*, eo5311. https://doi.org/10.2903/j.efsa.2018.5311

Azimonti, G., Galimberti, F., Marchetto, F., Menaballi, L., Ullucci, S., Pelliccioli, F., Caffi, A., Ceriani, L., Ippolito, A., Moretto, A., de Boer, W., & van der Voet, H. (2015). Comparison of NOEC values to EC10/EC20 values, including confidence intervals, in aquatic and terrestrial ecotoxicological risk assessment. *EFSA Supporting Publication* *12*, 274. https://doi.org/10.2903/sp.efsa.2015.EN-906

Brock, T. C. M., Hammers-Wirtz, M., Hommen, U., Preuss, T. G., Ratte, H. T., Roessink, I., Strauss, T., & van den Brink, P. J. (2015). The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environmental Science and Pollution Research*

International, *22*, 1160–1174. https://doi.org/10.1007/s11356-014-3398-2

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. R package version 1.3-0. https://CRAN.R-project.org/package=pwr.

Clarke, E., Sherrill-Mix, S., & Dawson, C. (2023). *ggbeeswarm: Categorical Scatter (Violin Point) Plots*. R package version 0.7.2. https://CRAN.R-project.org/package=ggbeeswarm.

Coady, K., Marino, T., Thomas, J., Currie, R., Hancock, G., Crofoot, J., McNalley, L., McFadden, L., Geter, D., & Klecka, G. (2010). Evaluation of the amphibian metamorphosis assay: Exposure to the goitrogen methimazole and the endogenous thyroid hormone L-thyroxine. *Environmental Toxicology and Chemistry*, *29*, 869–880.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Dang, Z. (2022). Amphibian toxicity testing for identification of thyroid disrupting chemicals. *Environmental Pollution*, *311*, 120006. https://doi.org/10.1016/j.envpol.2022.120006

Dardis, C. (2022). *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.6. https://CRAN.R-project.org/package=survMisc.

Eaton, A. (2025). *RMSTdesign: Sample size and power calculation for the difference in restricted mean survival time*. R package version 0.0.0.9000. https://anneae.github.io/RMSTdesign/.

Eaton, A., Therneau, T., & Le-Rademacher, J. (2020). Designing clinical trials with (restricted) mean survival time endpoint: practical considerations. *Clinical Trials*, *17*, 285–294. https://doi.org/10.1177/1740774520905563

European Chemicals Agency (ECHA)/European Food Safety Authority (EFSA) with the technical support of the Joint Research Centre (JRC); Andersson, N., Arena, M., Auteri, D., Barmaz, S., Grignard, E., Kienzler, A., Lepper, P., Lostia, A. M., Munn, S., Parra Morte, J. M., Pellizzato, F., Tarazona, J., Terron, A., & Van der Linden, S. (2018). Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA Journal*, *16*, 5311–5135. https://doi.org/10.2903/j.efsa.2018.5311.

European Commission. (2009). Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. *Official Journal of the European Union*, *309*, 1–50.

European Commission. (2018). Commission Regulation (EU) 2018/605 of 19 April 2018 amending Annex II to Regulation (EC) No 1107/2009 by setting out scientific criteria for the determination of endocrine disrupting properties. *Official Journal of the European Union*, *101*, 33–36.

European Food Safety Authority (EFSA) Panel on Plant Protection Products and their Residues (PPR). (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal*, *11*, 3290. https://doi.org/10.2903/j.efsa.2013.3290

Fort, D. J., Mathis, M. B., Pawlowski, S., Wolf, J. C., Peter, R., & Champ, S. (2017). Effect of triclosan on anuran development and growth in a larval amphibian growth and development assay. *Journal of Applied Toxicology: JAT*, *37*, 1182–1194. https://doi.org/10.1002/jat.3474.

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (3rd ed.). Sage.https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Green, J. W., Springer, T. A., & Holbech, H. (2018). *Statistical analysis of ecotoxicity studies*. John Wiley & Sons. https://doi.org/10.1002/9781119488798

Hartig, F. (2024). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.4.7. https://CRAN.R-project.org/package=DHARMa.

Haselman, J. T., Kosian, P. A., Korte, J. J., Olmstead, A. W., & Degitz, S. J. (2018). Effects of multiple life stage exposure to the fungicide prochloraz in *Xenopus laevis*: Manifestations of antiandrogenic and other modes of toxicity. *Aquatic Toxicology*, *199*, 240–251.

Haselman, J. T., Kosian, P. A., Korte, J. J., Olmstead, A. W., Iguchi, T., Johnson, R. D., & Degitz, S. J. (2016). Development of the larval amphibian growth and development assay: Effects of chronic 4-tert-octylphenol or 17β-trenbolone exposure in *Xenopus laevis* from embryo to juvenile. *Journal of Applied Toxicology: JAT*, *36*, 1639–1650. https://doi.org/10.1002/jat.3330.

Haselman, J. T., Olker, J. H., Kosian, P. A., Korte, J. J., Swintek, J. A., Denny, J. S., Nichols, J. W., Tietge, J. E., Hornung, M. W., & Degitz, S. J. (2020). Targeted pathway-based in vivo testing using thyroperoxidase inhibition to evaluate plasma thyroxine as a surrogate metric of metamorphic success in model amphibian *Xenopus laevis*. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *175*, 236–250.

Horikoshi, M., & Tang, T. (2016). *ggfortify: Data visualization tools for statistical analysis results*. https://CRAN.R-project.org/package=ggfortify

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift*, *50*, 346–363.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2024). *forecast: Forecasting functions for time series and linear models*. R package version 8.23.0. https://pkg.robjhyndman.com/forecast/

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*, 1–22. https://doi.org/10.18637/jss.v027.i03

Kassambara, A., Kosinski, M., & Biecek, P. (2024). *survminer: Drawing survival curves using 'ggplot2'*. R package version 0.5.0. https://CRAN.R-project.org/package=survminer

Mair, M. M., Kattwinkel, M., Jakoby, O., & Hartig, F. (2020). The minimum detectable difference (MDD) concept for establishing trust in nonsignificant results: A critical review. *Environmental Toxicology and Chemistry*, *39*, 2109–2123.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer

Nieuwkoop, P. D., & Faber, J. (1994). *Normal table of Xenopus laevis (Daudin): A systematical and chronological survey of the development from the fertilized egg till the end of metamorphosis*. Garland Publishing. https://doi.org/10.1201/9781003064565

Olker, J. H., Haselman, J. T., Kosian, P. A., Donnay, K. G., Korte, J. J., Blanksma, C., Hornung, M. W., & Degitz, S. J. (2018). Evaluating iodide recycling inhibition as a novel molecular initiating event for thyroid axis disruption in amphibians. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *166*, 318–331.

Organisation for Economic Co-operation and Development (OECD) (2006). *Current approaches in the statistical analysis of ecotoxicity data: A guidance to application, OECD series on testing and assessment, Number 54, ENV/JM/MONO(2006)18*. Environment Directorate, Organisation for Economic Co-Operation and Development. https://doi.org/10.1787/9789264085275-en

Organisation for Economic Co-operation and Development (OECD) (2007). *Final report of the validation of the Amphibian Metamorphosis Assay: Phase 2—Multi-chemical interlaboratory study*. OECD Environment, Health and Safety Publications. Series on Testing and Assessment No. 77. ENV/JM/MONO(2007)2. https://one.oecd.org/document/ENV/JM/MONO(2008)19/en/pdf

Organisation for Economic Co-operation and Development (OECD) (2009). *Test No. 231: Amphibian Metamorphosis Assay, OECD guidelines for the testing of chemicals, Section 2*. OECD Publishing. https://doi.org/10.1787/9789264076242-en

Organisation for Economic Co-operation and Development (OECD) (2015). *Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA), OECD guidelines for the testing of chemicals, Section 2*. OECD Publishing. https://doi.org/10.1787/9789264242340-en

Organisation for Economic Co-operation and Development (OECD) (2018). *Revised guidance document 150 on standardised test guidelines for evaluating chemicals for endocrine disruption, OECD series on testing and assessment, No. 150*. OECD Publishing. https://doi.org/10.1787/9789264304741-en

Organisation for Economic Co-operation and Development (OECD) (2019). *Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA), Section 2*. OECD Publishing. https://doi.org/10.1787/a13f80ee-en

Ortego, L., Olmstead, A., Banman, C., & Hall, T. (2014). *Improving the Amphibian Metamorphosis Assay: Experience with a time-to-stage design*. Proceedings, 35th SETAC North America Annual Meeting, Vancouver, BC, Canada, November 9–13, 2014.

Ortego, L. S., Olmstead, A. W., Weltje, L., Wheeler, J. R., Bone, A. J., Coady, K. K., Banman, C. S., Burden, N., & Lagadic, L. (2021). The Extended Amphibian Metamorphosis Assay: A thyroid-specific and less animal-intensive alternative to the larval amphibian growth and development assay. *Environmental Toxicology and Chemistry*, 40, 2135–2144.

Papadakis, M., Tsagris, M., & Fafalios, S. (2023). *Rfast: A collection of efficient and extremely fast R functions*. R package version 2.1.0. https://CRAN.R-project.org/package=Rfast

Pawlowski, S., Dammann, M., Weltje, L., Champ, S., Mathis, M., & Fort, D. (2019). Is normalized hindlimb length measurement in assessment of thyroid disruption in the amphibian metamorphosis assay relevant? *Journal of Applied Toxicology: JAT*, 39, 1164–1172.

Pedersen, T. (2024). *patchwork: The composer of plots*. R package version 1.3.0. https://CRAN.R-project.org/package=patchwork

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Royston, P., & Parmar, M. K. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13, 152–115.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39, 499–503. https://doi.org/10.2307/2531021

Seshan, V., & Whiting, K. (2023). *clinfun: Clinical trial design and data analysis functions*. R package version 1.1.1. https://CRAN.R-project.org/package=clinfun

Signorell, A. (2024). *DescTools: Tools for descriptive statistics*. R package version 0.99.57. https://CRAN.R-project.org/package=DescTools

Sjoberg, D., Baillie, M., Fruechtenicht, C., Haesendonckx, S., & Treis, T. (2024). *ggsurvfit: Flexible time-to-event figures*. R package version 1.1.0. https://CRAN.R-project.org/package=ggsurvfit

Szabo, A. (2024). *CorrBin: Nonparametrics with clustered binary and multinomial data*. R package version 1.6.2. https://CRAN.R-project.org/package=CorrBin

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified interface to visualize statistical result of popular R packages. *The R Journal*, 8, 474–489.

Therneau, T. M. (2024). *coxme: Mixed Effects Cox Models*. R package version 2.2-22. https://CRAN.R-project.org/package=coxme

Torchiano, M. (2020). effsize: Efficient effect size computation. R package version 0.8.1. https://CRAN.R-project.org/package=effsize. https://doi.org/10.5281/zenodo.1480624

Uno, H., Tian, L., Horiguchi, M., Cronin, A., Battioui, C., & Bell, J. (2022). *survRM2: Comparing restricted mean survival time*. R package version 1.0-4. https://CRAN.R-project.org/package=survRM2

U.S. Environmental Protection Agency (2009). *OPPTS 890.1100: Amphibian metamorphosis (frog)*. https://www.regulations.gov/document/EPA-HQ-OPPT-2009-0576-0002

U. S. Environmental Protection Agency (2013). *Scientific review of the Endocrine Disruptor Screening Program (EDSP); tier 1 screening assay and battery performance*. EPA-HQ-OPP-201300750023.

U.S. Environmental Protection Agency (2015). *OCSPP 890.2300. Larval Amphibian Growth and Development Assay (LAGDA)*. https://www.regulations.gov/document/EPA-HQ-OPPT-2009-0576-0018

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., & Bryan, J. (2022). *readxl: Read excel files*. R package version 1.4.1. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation*. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Girlich, M. (2022). *tidyr: Tidy messy data*. R package version 1.2.0. https://CRAN.R-project.org/package=tidyr

Wilke, C. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

Wood, J. (2024). *RcppAlgos: High performance tools for combinatorics and computational mathematics*. R package version 2.8.5. https://CRAN.R-project.org/package=RcppAlgos

© The Author(s) 2025. Published by Oxford University Press on behalf of the Society of Environmental Toxicology and Chemistry.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Environmental Toxicology and Chemistry, 2025, 44, 1435–1447
https://doi.org/10.1093/etojnl/vgaf067
Original Article