Critical evaluation of existing in vivo tests for endocrine modes of action

Main author: Natalie Burden (National Centre for the 3Rs (NC3Rs))

Co-authors: Constance Mitchell, Michelle Embry

INTRODUCTION

Some regional regulations now explicitly require that endocrine disrupting properties be investigated in the safety assessment process (e.g. EFSA/ECHA, 2018). This can include conducting new vertebrate tests for substances already on the market. Testing for activity usually involves in vitro and in vivo assays on estrogen, androgen, thyroid, and steroidogenesis pathways. For ecotoxicological evaluation, this can be performed across various species including mammals, amphibians and fish. Results that indicate activity usually trigger higher-tier in vivo assays to assess endocrine-relevant adverse effects over more extensive parts of the life cycle. Higher-tier assays are animal and resource-intensive and technically challenging to conduct. The aim of this project is to review the currently used in vivo test guidelines (TGs) to better understand: which criteria are fundamental to test performance (and where there are acceptable levels of flexibility); the biological relevance of the tests; and sources of variability, particularly for higher-tier studies. As well as assisting with interpretation issues, this activity will support evaluation of the utility of new approach methodologies.

METHODOLOGY

This cross-sector initiative is a collaborative effort, led by the Health and Environmental Sciences Institute (HESI) and the UK's NC3Rs, resulting from discussions held at an expert international workshop in early 2020. The biological and physico-chemical TG performance criteria (for both OECD and US EPA TGs) were extracted for the fish short-term reproduction assay (FSTRA), fish sexual development test (FSDT), medaka extended one generation reproduction test (MEOGRT), amphibian metamorphosis assay (AMA) and larval amphibian growth and development assay (LAGDA). Practical issues related to adhering to the criteria, and potential confounding effects on outcomes resulting from non-compliance with criteria were identified ('soft' information). Further, high-quality historical control data for 47 regulatory AMA studies were collated and data analysis was performed to assess compliance with TG performance and validity criteria.

RESULTS

Basic experimental designs are similar between all assays (e.g. no. of treatment groups/replicates). Physical factors vary between assays, reflecting the diversity of species and life stages specified. Biological performance criteria are assay- and species-specific. Achieving a pre-determined mean stage and animal weight (LAGDA, AMA) or reproductive output (FSTRA, FSTD, MEOGRT) in control treatments is deemed critical to the assay's success. Reproductive assays also specify hatching rates and the rate of survival of offspring in control treatments. Even where the infrastructure exists to meet challenging constant exposure conditions (i.e. for weeks or months), the biological complexity of the test subjects can be confounding. Inherent diversity within a test species, even in well-established laboratory strains, may result in test organisms that are smaller or larger than specified, reproduce at rates below minimum performance criteria, or develop in a way which does not meet the required performance criteria. For the specific AMA historical control data analyses, it was shown that routinely valid studies can be performed, but there needs to be some latitude in the interpretation of the TG criteria.

DISCUSSION

This review of the TGs currently used to assess endocrine-mediated activity/adversity has shown that, even while satisfying stringent performance criteria, these inherently complex studies may be sufficiently variable to limit the broader understanding of the biological effects of test chemicals. The availability of a comprehensive, high-quality AMA historical control dataset represents a unique opportunity to assess the performance of this important assay after OECD and US EPA validation. This dataset will be further explored to specifically investigate potential differences between laboratories and studies; relationships between endpoints and test conditions; sources of variability; power of the test design; histopathology data; and data interpretation. Findings from these analyses will inform the design of future tests in terms of reduction, replacement, and refinement of animal testing, support the interpretation of data, and form a knowledge base that could be used to improve future iterations of the TGs.