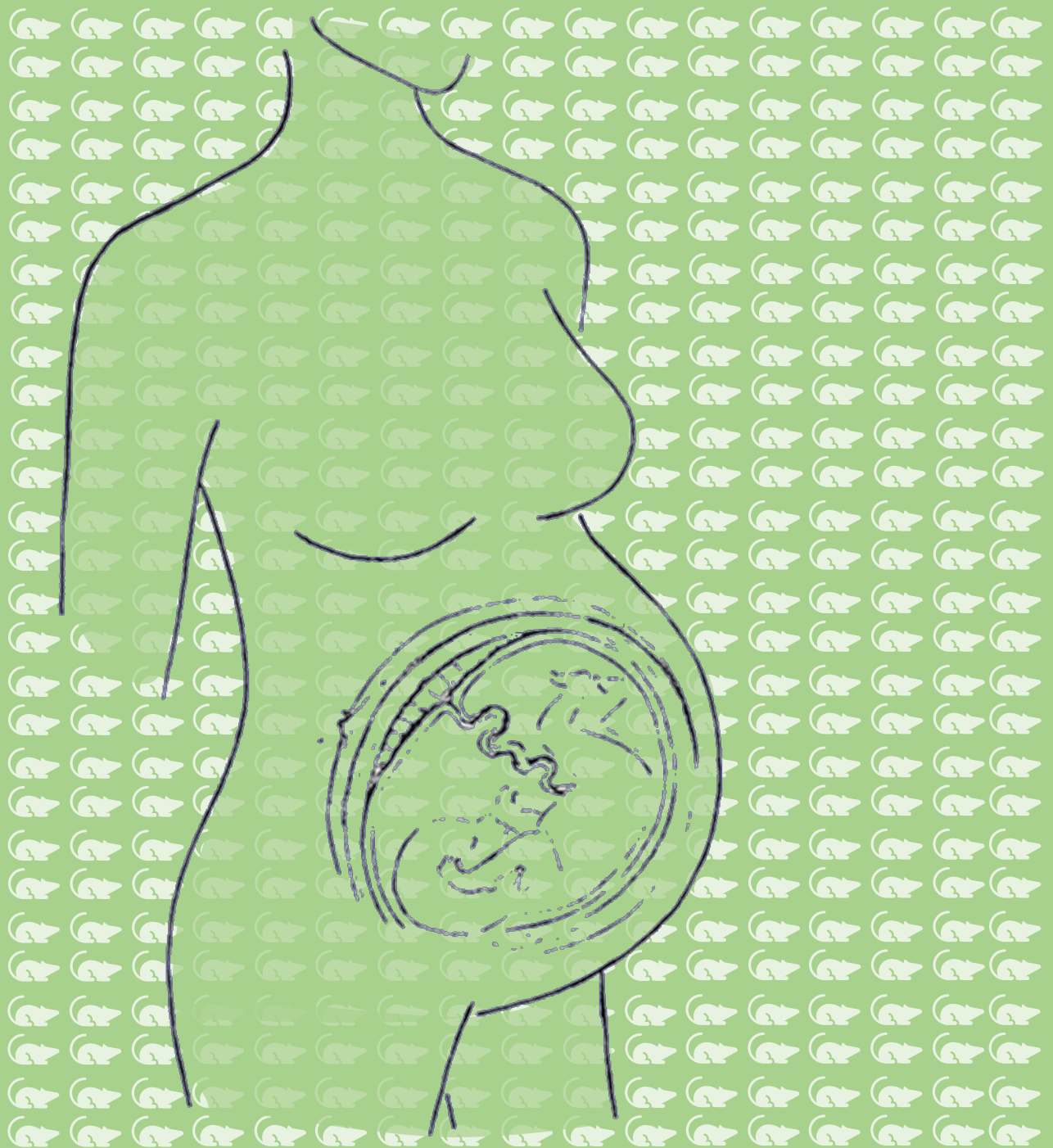
Developmental and Reproductive Toxicology (DART) Testing: The Next Generation

Working towards optimal implementation of the 3R principles in regulatory DART testing



Manon Beekhuijzen

DEVELOPMENTAL AND REPRODUCTIVE TOXICOLOGY (DART) TESTING: THE NEXT GENERATION

Working towards optimal implementation
of the 3R principles in regulatory DART testing

Manon Beekhuijzen

ISBN: 978-94-6233-974-3

Awesome cover design: Michiel Alexander de Raaf
Printed by: Gildeprint

Copyright © 2018 Manon Beekhuijzen

All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without prior permission of the author.

DEVELOPMENTAL AND REPRODUCTIVE TOXICOLOGY (DART) TESTING: THE NEXT GENERATION

Working towards optimal implementation
of the 3R principles in regulatory DART testing

Testen in ontwikkelings- en reproductietoxicologie: de volgende generatie

Werken naar een optimale invoering van de 3V principes in ontwikkelings- en reproductietoxicologie volgens de regelgeving

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 5 juli 2018 des ochtends te 10.30 uur

door

Maria Everdina Wilhelmina Beekhuijzen

geboren op 11 november 1976
te Valburg

Promotoren: Prof. dr. A.H. Piersma
Prof. dr. M. van den Berg

Dit proefschrift werd mogelijk gemaakt met financiële steun van Charles River Laboratories

This thesis is dedicated to my wonderful children, Bas en Roos, with all my love and affection

"If you are always trying to be normal, you will never know how amazing you can be"

CONTENTS

	Abbreviations	9
	Overview of regulatory guidelines	11
Chapter 1	General introduction.	13
Chapter 2	Effects of capillary microsampling on toxicological endpoints in juvenile rats.	35
Chapter 3	Satellite rats are redundant in embryo-fetal development studies.	53
Chapter 4	The underestimated value of OECD 421 and 422 repro screening studies: putting it in the right perspective.	71
Chapter 5	Update of OECD DART guidelines with endocrine disruptor relevant endpoints: Practical considerations.	87
Chapter 6	Implementing the extended one-generation reproductive toxicity study (EOGRTS): important points to consider.	105
Chapter 7	Guidance on the selection of cohorts for the extended one-generation reproduction toxicity study (OECD test guideline 443).	131
Chapter 8	From cutting edge to guideline: A first step in harmonization of the zebrafish embryotoxicity test (ZET) by describing the most optimal test conditions and morphology scoring system.	149
Chapter 9	Visualizing compound distribution during zebrafish embryo development: the effects of lipophilicity and DMSO.	177
Chapter 10	Summary, general discussion, future perspectives, and conclusion.	211
	Nederlandse samenvatting	231
	References	239
	Dankwoord	257
	List of publications	261
	Curriculum Vitae	271

ABBREVIATIONS

(Q)SAR	(quantitative) structure-activity relationship
3R	Replacement, Refinement and Reduction
ACSA	Agricultural Chemical Safety Assessment
AG	Anogenital
AGD	Anogenital distance
ALAT	Alanine aminotransferase
ALP	Alkaline Phosphatase
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AO	Adverse Outcome
AOP	Adverse Outcome Pathway
ASAT	Aspartate aminotransferase
B-cell	B lymphocyte
BMC	benchmark concentration
BMCt	benchmark concentration for teratogenicity
BMD	benchmark dose
BMR	benchmark response
BPR	Biocidal Products Regulation
CLP	Classification, labelling and packaging
CMS	Capillary Microsampling
CNS	Central nervous system
CRO	Contract Research Organization
DART	Developmental and Reproductive Toxicity
DBS	Dried Blood Spot
DES	diethylstilbestrol
DIT	developmental immunotoxicity
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNT	developmental neurotoxicity
DOHaD	Developmental Origins of Health and Disease
dpf	days post-fertilization
EC	European Commission
EC50	concentration with 50% effect
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
ECHA	European Chemicals Agency
EDC	Endocrine Disrupting Chemical
EDSP	Endocrine Disruptor Screening Program
EDTA	Ethylenediaminetetraacetic acid
EFD	Embryo-Fetal Development
EFSA	European Food Safety Authority
ELISA	Enzyme-Linked ImmunoSorbent Assay
ELS	early live stages
EMA	European Medicines Agency
EOGRTS	extended one-generation reproduction toxicity study
EPA	Environmental Protection Agency
EU	European Union
F0 or P	Parental generation
F1	Offspring of the F0 generation (parents).
F2	Offspring of the F1 generation.
FDA	Food and Drug Administration
FET	fish embryo acute toxicity
FIFRA	Federal Insecticide, Fungicide, and Rodenticide Act
G	Gauge
GD	Guidance Document
GLP	Good Laboratry Practice
GMP	Good Manufacturing Practice
GMS	general morphology score
h	Hour
HCD	Historical Control Data
HESI	Health and Environmental Science Institute
hEST	human Embryoic Stem cell Test
hpf	hours post-fertilization
IATA	Integrated Approaches to Testing and Assessment
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IgA	Immunoglobulin A
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IgM	Immunoglobulin M
ILSI	International Life Sciences Institute
Inorg. Phos.	Inorganic Phosphate
ITS	Integrated Testing Strategy
kDA	kiloDalton
KLH	keyhole limpet haemocyanin
KMD	kinetically-derived maximum dose
LABC	levator ani plus bulbocavernosus muscle complex
LC25	concentration with 25% mortality
LC50	concentration with 50% mortality
LC-MS	Liquid Chromatography- Mass Spectrometry
LD50	dose with 50% mortality
LOAEL	Lowest Observed Adverse Effect Level
logP	Logarithm of the partition coefficient
MAD	Mutual Acceptance of Data
MCH	Mean corpuscular hemoglobin

MCHC	Mean corpuscular hemoglobin concentration
MCV	Mean corpuscular volume
mEST	mouse Embryoic Stem cell Test
MIE	Molecular Initiating Event
MoA	Mode of Action
MOG	modified one generation
MTC	maximum tolerated concentration
MTD	maximum tolerated dose
N	Number
nAChR	nicotinic ACetylcholine Receptor
NC3R	National Centre for the Replacement, Refinement & Reduction of Animals in Research
NIH	National Institute of Health
NIH-ARAC	National Institute of Health Animal Research Advisory Committee
NK cell	Natural killer cell
NOAEL	No Observed Adverse Effect Level
NRC	National Research Council
NTP	National Toxicology Program
OECD	Organisation for Economic Co-operation and Development
OECD	Organisation for Economic Co-operation and Development
OPPTS	Office of Prevention, Pesticides and Toxic Substances
PFC	plaque-forming cells
PMN	Premanufacture Notice
PND	postnatal day
PP	Peroxisome Proliferators
PVS	perivitelline space
RDW	Red blood cell distribution width
REACH	Registration, Evaluation and Authorization of Chemicals
RMS	Rapporteur Member State
SD	Standard Deviation
SPF	Specific Pathogen Free
SRBC	Sheep red blood cell
STP	Society of Toxicologic Pathology
SVHC	Substance of Very High Concern
t/a	tonnes per annum (year)
T3	triiodothyronine (thyroid hormone)
T4	thyroxine (thyroid hormone)
T-cell	T lymphocyte
TDAR	T-cell dependant antibody response assay
TG	test guideline
TGE	total genistein equivalents
TI	teratogenic index
TK	Toxicokinetic
Tmax	Time at which the Cmax is observed
Tox21	Toxicology in the 21st Century
ToxCast	Toxicity Forecaster
TSCA	Toxic Substances Control Act
TSH	Thyroid-stimulating hormone
US	United States
v/v	volume/volume
VAMS	Volumetric Absorptive Microsampling
vs	versus
WBC	White blood cells
WEC	Whole Embryo Culture
WOCBP	Women Of Childbearing Potential
ZET	Zebrafish Embryotoxicity Test
ZF	Zebrafish

OVERVIEW OF REGULATORY GUIDELINES

Pharmaceuticals

ICH S5 (R2) Section 4.1.1.	Study of fertility and early embryonic development to implantation and Part II: Toxicity to Male Fertility
ICH S5 (R2) Section 4.1.2.	Study for effects on pre-and postnatal development, including maternal function
ICH S5 (R2) Section 4.1.3.	Study for effects on embryo-fetal development

Industrial chemicals

OECD 421	Reproduction/Developmental Toxicity Screening Test
OECD 422	Combined Repeated Dose Toxicity Study with the Reproduction/ Developmental Toxicity Screening Test
OECD 414	Prenatal Developmental Toxicity Study
OECD 416	Two generation reproduction toxicity study.
OECD 443	Extended One-Generation Reproductive Toxicity Study

Agrochemicals

OPPTS 870.3700	Prenatal Developmental Toxicity Study
OPPTS 870.3800	Reproduction and Fertility Effects

Chapter 1

General introduction

Adapted from:

Manon Beekhuijzen, The era of 3Rs implementation in Developmental and Reproductive Toxicity (DART) Testing: current overview and future perspectives, Reproductive Toxicology, 2017 Sep;72:86-96. doi: 10.1016/j.reprotox.2017.05.006. Epub 2017 May 25.

1.1. REPRODUCTION AND DEVELOPMENT

Reproduction (or breeding) is the biological process by which new individual organisms (offspring) are created by their parents. Reproduction is a fundamental characteristic of all known life; each individual organism exists due to reproduction. Sexual reproduction involves the combination of two gametes (which contain half the number of chromosomes of normal cells and are generated through meiosis), with a male fertilizing a female of the same species to produce a fertilized zygote. This creates offspring whose genetic traits are constructed from those of both parents.

Human development is the process of growing to maturity. In biological terms, this involves growth from a one-celled zygote to an adult human being. Fertilization occurs when the sperm successfully enters the ovum. The genetic material of the sperm and egg unite to form a single cell (i.e. zygote), and the germinal stage of prenatal development starts. The germinal stage is the time from fertilization, through the development of the early embryo, up until implantation. The zygote contains a complete set of genetic material and develops into the embryo. A week after fertilization the embryo sticks to the lining of the uterus. This causes a decidual reaction in which the uterine cells multiply and enclose the embryo. In the meantime, the embryo proliferates and changes into embryonic and extra-embryonic tissue, the latter creating the fetal membranes and the placenta. Embryogenesis is complete after closure of the palate around 8 weeks of gestation, as all major structural features of the embryo are formed. This is followed by fetal growth and development.

The steps involved in the human reproductive and developmental cycle are shown in Figure 1.1.

1.2. REPRODUCTIVE TOXICOLOGY

Reproductive toxicity includes adverse effects on sexual function and fertility in adult males and/or females, as well as developmental toxicity in the offspring. The distinction between developmental and reproductive toxicity is somewhat arbitrary in that developmental exposures can result in effects on reproduction, and vice versa [OECD GD 43, 2008]. Adverse effects on reproductive ability or capacity include alterations in the female and male reproductive organs or related endocrine systems i.e., effects on onset of puberty, gamete production and transport, reproductive cycle normality, sexual behaviour, fertility, parturition, premature reproductive senescence, or modifications in other functions that are dependent on the integrity of the reproductive systems (e.g. lactation) [OECD GD 43, 2008].

One well-known example of a chemical causing reproductive toxicity is diethylstilbestrol (DES). This was developed as an orally active estrogen prescribed to pregnant women to prevent miscarriages. Long-term effects of DES have been recorded in the mothers exposed to DES and their offspring (DES-daughters and DES-sons). Adverse pregnancy outcomes, infertility, cancer, and early meno-

pause have been discovered in women exposed to DES, and some events occur in their offspring and subsequent generations. Historically, DES was considered to cause its effects by a mutational event. However, it was never proven that DES changes the gene structure. Currently, this theory is outdated as it is thought that DES causes its effects by epigenetic alterations. [Al Jishi, 2017].

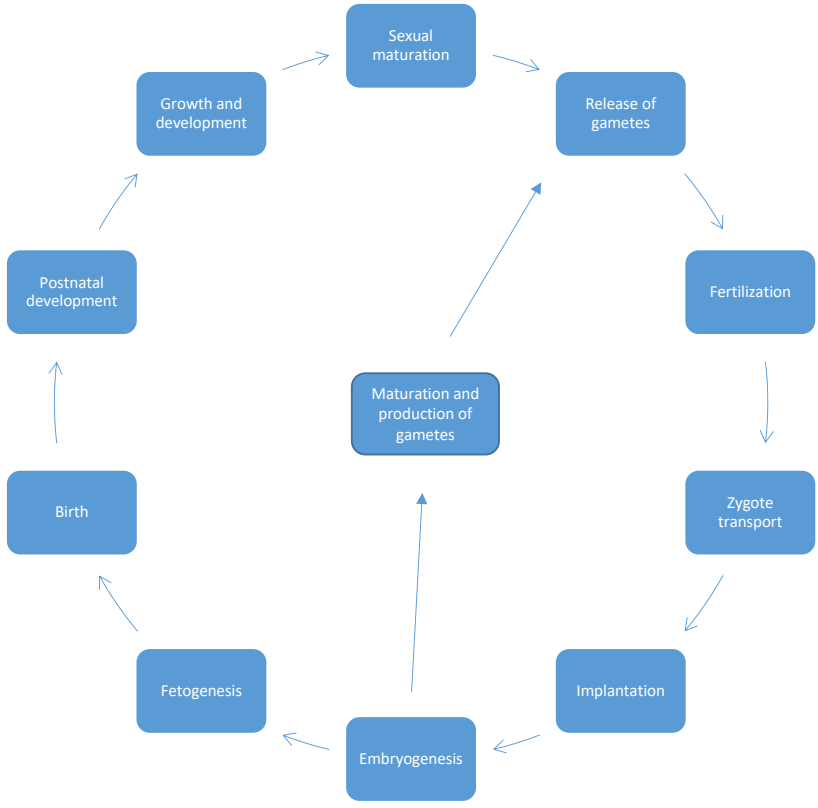


Figure 1.1. The human reproductive and developmental cycle.

Developmental toxicity is defined as the occurrence of adverse effects on the developing organism that may result from exposure prior to conception (either parent), during prenatal development, or postnatally to the time of sexual maturation. Adverse developmental effects may be detected at any point in the life span of the organism. The major manifestations of developmental toxicity include death of the developing organism, structural abnormality, altered growth, and functional deficiency [Selevan, 2000]. Developmental exposures may result in health effects observed prenatally and at birth (such as spontaneous abortion, stillbirth, low birth weight, small for gestational age, infant mortality, and malformations), in childhood (effects such as asthma, cancer, neurologic and behavioral effects), and as adults and into old age (such as cancer, heart disease, and degenerative neurologic/ behavioral disorders) [Selevan, 2000].

The thalidomide tragedy of the early 1960s increased our understanding of developmental toxicology, providing a clear example of an agent that produced minimal toxicity to adults but a high degree of embryotoxicity [Ujházy, 2012]. Maternal exposure to the mild sedative-hypnotic agent thalidomide, used for morning sickness during pregnancy, caused characteristic reduction deformities of the limbs, ranging from hypoplasia of one or more digits to the total absence of all limbs. An example of the thalidomide embryopathy is phocomelia (i.e. the structures of the hand and feet may be reduced to a single small digit, or may appear virtually normal but protrude directly from the trunk). The proof was presented by the independent discoveries of Lenz [Lenz, 1961] and McBride [McBride, 1961], which led to the worldwide interest in teratology.

Whether this disaster could have been prevented remains unclear. The packaging and information of contraindications of drugs was minimal when compared to the way drugs are sold today. More importantly, the state of drug testing in the 1950s and 1960s was different when compared to today's standards. According to Grunenthal (the pharmaceutical company that marketed thalidomide), the testing of thalidomide was carried out using the standards of the time, however it remains a matter of debate as to precisely what testing was carried out [Vargesson, 2015].

The aim of experimental teratology in the post-thalidomide period has been the exact explanation of causes and mechanisms of the rise of congenital developmental disorders. Several factors that determine the teratogenicity of an exposure have been set forth as the principles of teratology by Wilson [Wilson, 1977].

He formulated a concept of six main principles of teratology that are generally accepted to this day [Friedman, 2010] and consist of:

1. Susceptibility to teratogenesis depends on the genotype of the conceptus and the manner in which this interacts with environmental factors.
2. Susceptibility to teratogenic agents varies with the developmental stage at the time of exposure.
3. Teratogenic agents act in specific ways (mechanisms) on developing cells and tissues to initiate abnormal embryogenesis (pathogenesis).
4. The final manifestations of abnormal development are death, malformation, growth retardation, and functional disorder.
5. The access of adverse environmental influences to developing tissues depends on the nature of the influences (agent).
6. Manifestations of deviant development increase in degree as dosage increases from the no-effect to the totally lethal level.

One of these principles states that susceptibility to teratogenesis varies with the developmental stage at the time of exposure, and one principle defines that manifestations of abnormal development depend on dose and duration of a teratogenic exposure. These indicate that not all exposures deemed as teratogenic are actually teratogenic all the time; the timing and dose of a partic-

ular exposure during pregnancy often determine the kind and extent of its teratogenic potential. The human embryonic period, during which organogenesis takes place, occurs between implantation at around 14 days to around 60 days post-conception. This is usually the most sensitive period to teratogenesis when exposure to a teratogenic agent has the greatest likelihood of producing a malformation. For example, administration of many established major teratogenic drugs, such as isotretinoin, valproic acid, warfarin, or high-dose methotrexate, in specific gestational windows in the first trimester is associated with a high risk of having a baby with a congenital malformation, but the risk significantly decreases in the second or third trimesters of pregnancy [Alwan, 2015].

The principle of critical windows of sensitivity is depicted in Figure 1.2. This shows a time line for human development in which highly sensitive windows for morphological abnormalities are identified, and these are compared to less sensitive times that are more likely to be associated with physiological defects and minor morphological abnormalities [Selevan, 2000]. Later, an adaptation (Figure 1.3) of these efforts was made for commonly studied adverse outcomes for humans, broader than malformations and physiological defects, but still limited to preconceptional and prenatal exposures [Selevan, 2000].

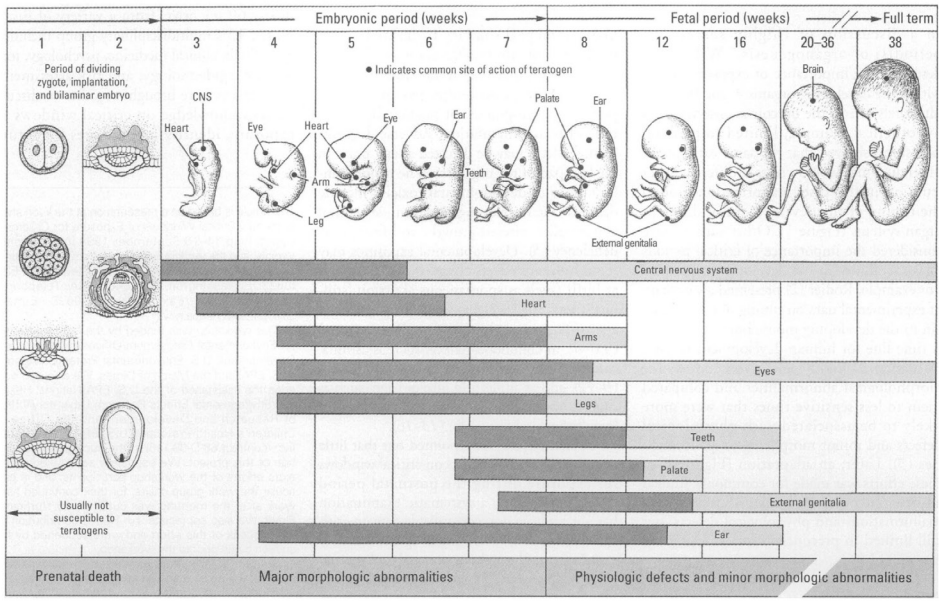


Figure 1.2. Schematic illustration of the sensitive or critical periods in human development. Dark gray denotes highly sensitive periods; light gray indicates stages that are less sensitive to teratogens [copied from Selevan, 2000].

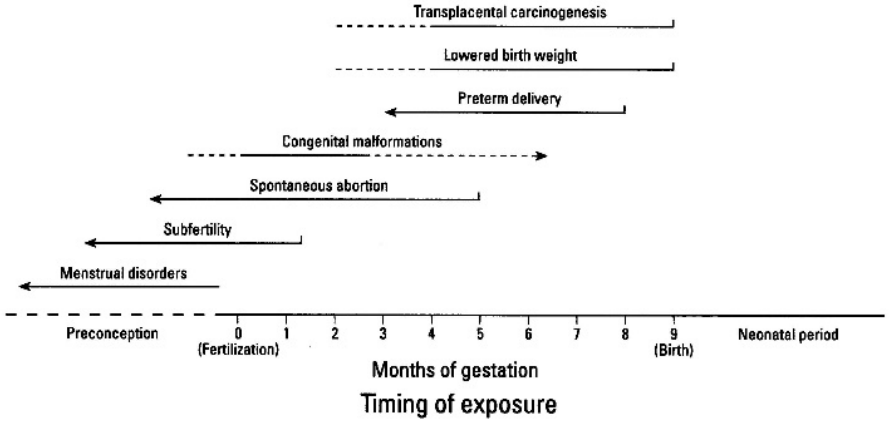


Figure 1.3. Reproductive outcomes associated with timing of maternal exposure. Solid lines indicate the most probable timing of exposure for a particular outcome; dotted lines indicate less probable but still possible timing of exposure. Arrows suggest that a defined cutoff point for exposure to a specific outcome is not known [copied from Selevan, 2000].

Moreover, reproductive toxicants have a threshold of adversity. This is in contrast to genotoxic carcinogens which are considered to have "no safe exposure level". The idea of no exposure level without adverse effects stems from the principle that even exposure to a single molecule of a genotoxic chemical may cause a mutation in the DNA, which will transform the affected cell into a neoplastic cell that escapes normal cell cycle homeostasis, resulting in uncontrolled cell proliferation and carcinogenesis [Piersma, 2011]. To this complexity of the threshold of adversity, reproductive toxicity adds the factor of the vulnerable window. The life cycle represents a range of species appearances (e.g., the sperm or egg, the embryo, the newborn, or the adult). Each of these life cycle stages have their own characteristics and homeostatic mechanisms and therefore they also have their specific vulnerabilities to chemical exposures, and based on ample data there is considerable evidence for the presence of dose levels with no measurable increased risk [Piersma, 2011].

The thalidomide tragedy triggered the development of more structured regulations and implementation of guidelines for Developmental and Reproductive Toxicity (DART) testing.

1.3. REGULATORY DART TESTING

For hazard assessment of chemicals in the regulatory domain, animal studies are being performed as default. For DART testing, the first harmonized guidelines were published in the 1980's for industrial chemicals (OECD 414 for developmental toxicity in 1981, OECD 415 and 416 for reproductive toxicity in 1983) [OECD 414, OECD 415, OECD 416] and in the 1990's for pharmaceuticals (ICH S5 for DART in 1993) [ICH S5]. In addition to these definitive tests, a relatively quick screening method which can give initial clues about possible fertility and developmental effects of chemicals is the OECD 421 screening study adopted in 1995, and the combination with 28-day repeated dose, OECD 422, adopted in 1996 [OECD 421, OECD 422]. The most recent adopted guideline is OECD 443, the extended one-generation reproductive toxicity study (EOGRTS), which was finalized in 2011 [OECD 443] and was developed as replacement for the two-generation reproduction toxicity study [OECD 416].

Tests for reproductive toxicity should ideally be able to identify one or more of the following effects on reproduction [OECD GD 43, 2008]:

- Impairment of male or female reproductive functions or capacity, i.e., adverse effects on libido, oestrus cycle, sexual behavior, any aspect of spermatogenesis or oogenesis, or hormonal activity or physiological response which would interfere with the capacity to fertilize, fertilization itself or development of the fertilized ovum up to and including implantation.
- Induction of non-inheritable harmful effects on the progeny, i.e., in the widest sense, any effect interfering with normal development, both before and after birth should be included. Morphological malformation(s) and functional disturbances (e.g., hormonal reproductive effects, neurological) should be evaluated.

For DART evaluation, two broad categories of toxicity are considered: reproductive and developmental toxicity. In the reproductive toxicity category, three classes of toxicity are evaluated: toxic effects on fertility, parturition, and lactation. In the developmental toxicity category, four classes of toxicity are evaluated: mortality, dysmorphogenesis (structural alterations), alterations to growth, and functional toxicities.

Despite the variety of compounds being tested (foods, drugs, agrochemicals, industrial chemicals, etc.), general designs for guidelines are remarkably similar. Table 1.1 provides an overview of the most recent versions of DART guidelines for pharmaceuticals, industrial chemicals and agrochemicals, with details on the required numbers of animals and performance timing. The guidelines vary primarily in the duration of dosing (which attempts to anticipate or simulate human exposure) and in the intent of the study. Table 1.1 shows that the DART study designs for industrial chemicals and agrochemicals are similar, but for pharmaceuticals the study designs are different. This difference is due to controlled use of drugs (i.e. known dosage and duration) versus uncontrolled exposure to industrial- and agrochemicals.

According to Commission Regulation (EU) No. 283/2013, DART testing for agrochemicals should include a two-generation study in the rat and developmental toxicity study in the rat and rabbit [Commission Regulation (EU) No. 283/2013].

Under REACH, data requirements for industrial chemicals depend on the volume of a substance that will be produced or imported into the European Union, using annual tonnage as a proxy for the likelihood of or importance of human exposure [ECHA R.7a, 2016]. Substances produced or imported at less than 1 metric tonne (1,000 kg) do not require registration under REACH. The registration begins at 1 tonne/year and the extent of toxicological information to be provided increases with tonnage in different tonnage bands as specified in the REACH Annexes. Testing for reproductive and developmental toxicity begins at 10 tonnes/year. A screening reproductive/developmental test (OECD 421 or 422) in the rat is required at 10 tonnes/year and above. A prenatal developmental toxicity test (OECD Test Guideline 414) in one species (usually the rat) is required at 100 tonnes/year and above. An extended one-generation reproductive toxicity study (EOGRTS; OECD 443) in the rat and a prenatal developmental toxicity study in a second species (usually the rabbit) is required at 1,000 tonnes/year.

DART guideline ICH S5 for pharmaceuticals mentions that it is perceived that tests in which animals are treated during defined stages of reproduction better reflect human exposure to medicinal products and allow more specific identification of stages at risk [ICH, 2005]. According to this guideline, one complete life cycle (i.e. from conception in one generation through conception in the following generation) can be subdivided into the following stages:

A) Premating to conception (adult male and female reproductive functions, development and maturation of gametes, mating behavior, fertilization).
B) Conception to implantation (adult female reproductive functions, preimplantation development, implantation).
C) Implantation to closure of the hard palate (adult female reproductive functions, embryonic development, major organ formation).
D) Closure of the hard palate to the end of pregnancy (adult female reproductive functions, fetal development and growth, organ development and growth).
E) Birth to weaning (adult female reproductive functions, neonate adaptation to extrauterine life, preweaning development and growth).
F) Weaning to sexual maturity (postweaning development and growth, adaptation to independent life, attainment of full sexual function).

The most probable option for DART testing of drugs is a combination of studies for effects on:
- Fertility and early embryonic development in the rat (stages A to B)
- Embryo-fetal development in the rat and rabbit (stages C to D)
- Pre- and postnatal development, including maternal function in the rat (stages C to F)

Table 1.1 (discussed on page 24). Overview of current DART guidelines for pharmaceuticals, industrial chemicals and agrochemicals, numbers of animals used and trigger points when needed.

GUIDELINES	WHEN NEEDED	NUMBER OF ANIMALS[1]
Pharmaceuticals[2, 3]		
ICH S5 (R2) Section 4.1.1. Study of fertility and early embryonic development to implantation and Part II: Toxicity to Male Fertility	Should be completed before the initiation of large scale or long duration clinical trials (e.g. phase 3 trials), since an evaluation of the reproductive organs is performed in the repeated-dose toxicity studies.	F_0-generation: 80 males and 80 females F_1-generation: approximately 960 fetuses
ICH S5 (R2) Section 4.1.3. Study for effects on embryo-fetal development in rats	In the EU and Japan definitive developmental toxicity studies in two species are required for inclusion of women of childbearing potential (WOCBP) in any clinical study. However, the document defines situations where clinical trials with WOCBP may progress without the 2 developmental toxicity studies in these regions. One circumstance could be a short-duration (e.g., 2 weeks) clinical study with precautions taken to prevent pregnancy. Another circumstance is where the clinical study must include WOCBP since the disease is prominent in women, but sufficient precautions are used to prevent pregnancy. In the relevant regions, the guidance also allows for the inclusion of WOCBP (up to 150) using adequate birth control methods for a relatively short duration (up to 3 months) if preliminary developmental toxicity data are available from two species. In the United States, the assessments of embryo-fetal development can be deferred until before phase 3 clinical studies for WOCBP using precautions to prevent pregnancy.	F_0-generation: 88 females F_1-generation: approximately 1056 fetuses
ICH S5 (R2) Section 4.1.3. Study for effects on embryo-fetal development in rabbits	See above.	F_0-generation: 88 females F_1-generation: approximately 792 fetuses
ICH S5 (R2) Section 4.1.2. Study for effects on pre-and postnatal development, including maternal function	Should be submitted for marketing approval.	F_0-generation: 96 females F_1-generation: approximately 1152 pups
Industrial chemicals [4]		
OECD 421, Reproduction/Developmental Toxicity Screening Test (2016)	REACH Annex VIII (applicable for any registration of 10 tonnes or more per year)	F_0-generation: 48 females and 40 males F_1-generation: approximately 480 pups
OECD 422, Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test (2016)	REACH Annex VIII	F_0-generation: 48 females and 40 males F_1-generation: a pproximately 480 pups

OECD 414, Prenatal Developmental Toxicity Study in rats (2001)	REACH Annex IX (applicable for any registration of 100 tonnes or more per year)	F_0-generation: 88 females F_1-generation: approximately 1056 fetuses
OECD 414, Prenatal Developmental Toxicity Study in rabbits (2001)	REACH Annex X (applicable for any registration of 1000 tonnes or more per year)	F_0-generation: 88 females F_1-generation: approximately 792 fetuses
OECD 443, Extended One-Generation Reproductive Toxicity Study (2012)	REACH Annex X	F_0-generation: 100 females and 100 males F_1-generation: approximately 1200 pups
Agrochemicals [5]		
OPPTS 870.3700, Prenatal Developmental Toxicity Study in rats (1998)	Before registration	F_0-generation: 88 females F_1-generation: approximately 1056 fetuses
OPPTS 870.3700, Prenatal Developmental Toxicity Study in rabbits (1998)	Before registration	F_0-generation: 88 females F_1-generation: approximately 792 fetuses
OPPTS 870.3800, Reproduction and Fertility Effects (1998)	Before registration	F_0-generation: 96 females and 96 males F_1-generation: approximately 1152 pups F_2-generation: approximately 1152 pups

[1] Assuming litter size of 12 rats/litter and 9 rabbits/litter. Excluding satellite animals (for toxicokinetic assessment).

[2] ICH S5 (R2): International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline on Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility, parent guideline dated 24 June 1993 (addendum dated 9 November 2000 incorporated in November 2005).

[3] Timing is described in ICH M3(R2).

[4] Timing is described in ECHA. Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7a. Endpoint Specific Guidance, Version 5.0; December 2016.

[5] Timing is described Commission Regulation (EU) No 283/2013 of 1 March 2013, setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. Official Journal of the European Union L 93, 1-84, 2013.

1.4. NUMBER OF ANIMALS USED IN REGULATORY DART TESTING

All regulatory DART tests demand large numbers of animals, as parental and offspring generations are involved, and depending on circumstances, developmental toxicity studies can be required to be executed in two species (i.e. one rodent and one non-rodent). Since adoption of the first version of these guidelines, many years passed without major updates. However, in the last couple of years, significant changes in these guidelines have been or are being implemented. These changes have in part been guided by the ethical drive to reduce, refine and replace (3R) animal testing, as well as the addition of endocrine disruptor relevant endpoints. Recent applied regulatory changes have focused on reduction and refinement, and not yet on replacement.

Table 1.1 provides an overview of the most recent versions for DART guidelines with details on the required numbers of animals and performance timing. This table shows that a total of approximately 4000 to 4500 of animals are needed for each type of compound to fulfill a complete DART package (as described in the former paragraph).

1.5. SIGNIFICANCE OF EXPERIMENTAL DATA AND THEIR RELEVANCE TO HUMANS

Throughout many years, regulatory DART testing has proven its usefulness, as no major issues like those with DES and thalidomide (marketed before the implementation of modern hazard identifications) occurred. However, at the same time, inadequacies of animal safety testing have become evident. Uncertainties came up about the relevance of animal findings for the situation in man [Piersma, 2014].

For example, aspirin is teratogenic in rodents but not in humans [Sullivan, 1999], and Thalidomide-induced phocomelia is seen after in utero exposure in humans and rabbits, but not in rodents [Ito, 2012]. In addition, sulfoxaflor-induced fetal abnormalities and neonatal death in rats occur via a single MoA comprising sustained activation of the rat fetal-type muscle nicotinic acetylcholine receptor resulting in a sustained muscle contracture. This MoA is considered not relevant for humans, given fundamental qualitative differences in sulfoxaflor agonism on the rat versus the human muscle nAChR [Ellis-Hutchings, 2014]. Another example are a variety of substances such as hypolipidemic drugs, phthalate ester plasticizers, pesticides, and industrial solvents, which have shown to increase the size and number of peroxisomes in rats and mice. These are called peroxisome proliferators (PP) and several of them have shown to be carcinogenic, inducing tumors upon chronic administration to rats and/or mice. However, the human relevance of rodent tumors induced by PP has been a subject of debate over the last decade. Based on a human relevance mode of action (MoA) framework it was concluded that despite the presence of similar pathways in humans, it is unlikely that the proposed MoA for rodent tumors is plausible in humans, taking into account kinetic and dynamic factors [Lai, 2004]. The same applies to long-term hypothyroid-

ism that causes thyroid tumor formation in adult rodents, for which it is widely recognized that the rodent MoA is not relevant for humans because of a decreased sensitivity of the human thyroid pituitary axis [Meek, 2003].

Likewise, the temporal equivalence factors must be taken into account. There are 10- to 100-fold differences in time span for pre- and postnatal development up to sexual maturity, in rodents compared with humans. The period of major organ development in the rat covers a period of about 10 days, which compares with about 56 days in humans. Development of the limbs for example takes 3 days in the rat compared to 3 weeks in humans. This can be contrasted with the relatively small pharmacokinetic differences, perhaps of 2- or 3-fold in plasma levels of chemicals with the same administered dose in the rat and human. The consequence of these differences is that single episodes of high chemical exposure have greater opportunity to produce permanent effects on development in rodents compared with humans. The limited evidence from acute poisonings in pregnant women tends to support this hypothesis [Sullivan, 1999].

Furthermore, through the years additional parameters of toxicity have been identified that could be included in testing and may improve hazard assessment. These comprise of endocrine disruption, developmental neurotoxicity and developmental immunotoxicity.

There has been considerable discussion in both the scientific literature and the lay press regarding the possibility that environmental chemicals, through their effects on endocrine function, are responsible for a number of reproductive and developmental anomalies in a wide range of wildlife species from invertebrates, through fish, reptiles, birds and mammals, including humans [Cooper, 1997]. During early development (i.e. the fetus and infant) exposures to proper levels of hormones are needed for normal ontogenic processes to occur. Too much or too little of a hormone can result in neurological impairments (eg, thyroid hormone deficiency), disorders of sexual development (eg, abnormal androgen or estrogen levels), or even death [Gore, 2015]. Exposure to exogenous compounds such as endocrine-disrupting chemicals (EDCs) that can mimic or antagonize these and other hormonal systems change the developmental trajectory. Depending on when an individual is exposed, the nature of the EDC, and its levels, an adverse outcome may be evident at birth, or there may be a latent disease outcome. This latter concept, often referred to as the developmental origins of health and disease (DOHaD), underlies latent effects of EDCs on every endocrine system studied to date [Gore, 2015]. Moreover, the central nervous system, endocrine and immune systems communicate through multiple anatomical and hormonal-neuropeptide routes. Controlled interactions between these systems are believed to be critical for the maintenance of a homeostatic balance within the body and good health [Taub, 2008]. An imbalance or alterations in these systems in response to disease, stress, injury and/or metabolic alterations can lead to significant changes. The lines of communication between the immune system and these various neuronal and endocrine organ systems constitute specific axes of interactions, which have been shown to have a profound impact on immune function, disease development and susceptibility to infections and disease [Taub, 2008]. A large body of research in both animals and humans

documents the permanent and definitely adverse consequences of developmental disruption of thyroid function. Even transient disruption of normal thyroid homeostasis will lead to adverse outcomes in the developing nervous system, including brain development and cognitive changes [Crofton, 2007].

Based on the increasing incidences of adverse outcomes associated with these endpoints [Kortenkamp, 2011] and with the aim to improve human hazard assessment, evaluation of endocrine disruption, neurotoxicity and immunotoxicity has recently been added to regulatory DART guidelines. All three endpoints were added to the EOGRTS (OECD 443). In addition, endocrine parameters were added to the Repeated Dose 28-Day Oral Toxicity Study in Rodents (OECD 407), the Reproduction/Developmental Toxicity Screening Test (OECD 421) and the Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test (OECD 422). It is anticipated that the Prenatal Developmental Toxicity Study (OECD 414) and the Repeated Dose 90-Day Oral Toxicity Study in Rodents (OECD 408) will shortly be updated with endocrine endpoints.

1.6. THE 3R PRINCIPLES

The guiding principles supporting humane and ethical use of animals in scientific research are called the three Rs (i.e. replace, reduce and refine) and were first described by Russell and Burch in 1959 [Russell, 1959]. Any researcher planning to use animals in their research must first show why there is no alternative and what will be done to minimize animal numbers and suffering, i.e.: 1) Replace the use of animals with alternative techniques, avoiding the use of animals altogether, 2) Reduce the number of animals used to a minimum, to obtain sufficient information from fewer animals, and 3) Refine the way experiments are carried out, to make sure animals suffer as little as possible [Russell, 1959]. This includes better housing and improvements to procedures which minimize pain and suffering and/or improve animal welfare. Important to realize is that these improvements can often easily be incorporated without losing the feasibility to answer the scientific question. In many examples, this would even improve the scientific outcome.

The 3Rs are incorporated as an important notion for humane use of animals in research into several legislations, for instance in the European Union (EU), Brazil and Japan, and they are embedded in the respective Animal Welfare Acts in the United States, in China and in India [Törnqvist, 2014]. The principles of the 3Rs are also incorporated in international guidelines for toxicity testing, such as the guidelines developed by the OECD, and into regulatory guidelines used for toxicity evaluations for chemicals, agrochemicals, cosmetics and pharmaceuticals [Törnqvist, 2014]. Important to know is that the EU directive on the protection of animals used for scientific purposes adopted in 2010 demands of EU members to increase 3R collaboration and asks to establish national 3R centers [EU Directive 2010/63/EU, 2010].

Looking at the high animal usage in DART testing (paragraph 1.4), questionable relevance of animal testing for human risk assessment (paragraph 1.5) and importance of the 3R principles (paragraph 1.6), a lot of gain can be made in this field of regulatory DART research by improvement of currently used testing methods.

Since several decades, extensive research has been done into alternative test methods. This regards both refined in vivo testing as well as the search for animal-free alternative (in silico and in vitro) methods. The objectives of this research are three-fold: 1) implementation of the 3R principles, 2) reduce cost and testing time, and 3) improve human relevance/extrapolation by achieving a more mechanistic understanding of the nature of toxicity at the molecular, cellular, and organ level. Additionally, robust alternative assays based on human biological material would circumvent the need for interspecies extrapolation [Piersma, 2014].

As described by the National Research Council (NRC) report "*Toxicity testing is approaching scientific pivot point. It is poised to take advantage of the revolutions in biology and biotechnology*". Advances in toxicogenomics, bioinformatics, systems biology, epigenetics, and computational toxicology could transform toxicity testing from a system based on whole-animal testing to one founded primarily on in vitro methods that evaluate changes in biologic processes using cells, cell lines, or cellular components, preferably of human origin [NRC, 2007]. But for DART testing it will need many years before these changes might be implemented in the regulatory arena as currently no validated DART in vitro assay is implemented by regulatory agencies.

Recently, this has been done for 'more simple endpoints' (i.e. single cell or tissue type) like skin sensitization, which was triggered by the total animal testing ban in cosmetics which came into force with the Cosmetic Regulation n°1223/2009 [EU Regulation (EC) No 1223/2009, 2009]. The testing ban on finished cosmetic products applies since September 2004 and for ingredients or combination of ingredients since March 2009, with the exception of repeated-dose toxicity, reproductive toxicity, and toxicokinetics. For these specific assessments, the marketing ban applies since March 2013, irrespective of the availability of alternative non-animal tests. Recently, three in vitro methods for assessment of skin sensitization have been validated [Clouet, 2017] and cover specific key events within the skin sensitization Adverse Outcome Pathway (AOP) as outlined by the OECD [OECD, 2012].

This paradigm shift towards greater use of non-animal models has been recognized in a number of initiatives aimed at improving pharmaceutical development and environmental protection [Holmes, 2010]. The US Food and Drug Administration [FDA, 2004], European Innovative Medicines Initiative [Anon, 2006] and US National Research Council Committee on Toxicity and Assessment of Environmental Agents [NRC, 2007], have acknowledged the limitations of animal models as a major bottleneck in the development and assessment of the efficacy and safety of medicines and assuring protection from the potential adverse health effects of chemicals present in the environment. These initiatives/reports recognize current in vivo approaches do not fully meet today's

complex demands for toxicity testing and that there is a need for non-animal alternative models to be incorporated into safety assessment strategies. This has been taken up by a working group to update the ICH S5 guideline.

Russell and Burch saw replacement as the ultimate goal for laboratory animal based research, education and testing, with the other two, reduction and refinement, being more readily achievable in the short term [Russell, 1959]. This is certainly applicable for DART testing as regulatory improvements have been made in respect of refinement and reduction, but replacement of current animal testing by alternative methods is still under development.

In DART testing one relatively simple adjustment in the method of blood sampling can make a great improvement on refinement and ultimately reduction of animal use. Blood is routinely sampled from laboratory animals in biomedical research, and many of the commonly applied sampling techniques require anesthesia [Arnold, 2010]. Awakening from anesthesia is considered very stressful and can easily be avoided when using refined blood sampling techniques (e.g. jugular vein blood sampling).

In addition, in rats, satellite groups have been used for exposure assessment to isolate toxicokinetic (TK) evaluation from the main study toxicity evaluations such that blood sampling does not induce artifacts and hence interfere with toxicity evaluations. This was in part based on the large blood volumes previously required for bioanalysis of the drug. Improved small molecule bioanalytical sensitivity and concomitant decreased sample volume requirements provide an opportunity to reconsider how TK data is collected in rat toxicity studies, most prominently evaluating TK in main study animals and reducing or even eliminating satellite groups. Especially in juvenile studies this would save a substantial number of animals. Advances in bioanalytical sensitivity have enabled smaller blood collection requirements, most notably “microsampling.” Microsampling is most commonly defined as individual blood samples that are ≤ 50 µL, in contrast with more conventional samples of 200 µL or greater [Chapman, 2014a; Chapman, 2014b]. Over the last decade, multiple methods, devices, and techniques utilizing microsampling have been described to reduce the number of animals required for toxicity testing. Examples include dried blood spot (DBS), capillary microsampling (CMS), and volumetric absorptive microsampling devices (VAMS) [Harstad, 2016]. All of these approaches require changes either to sample collection techniques, animal handling, bioanalytical sample preparation, and/or study designs. For many, there is no single best way or best technique as long as the final choice is technically feasible, reproducible, and adheres to ethical standards of animal welfare. The ultimate focus should remain on the reduction of animals utilized to answer the intended question or regulatory requirement.

Another way to markedly reduce the number of animals in DART testing is to omit the second generation in the multi-generation reproduction test for regulatory chemical assessment, which could save at least 1200 rats per study. The paper by Cooper et al in 2006 formed the basis [Cooper, 2006] of the in 2011 adopted Extended One-Generation Reproductive Toxicity Study (EOGRTS)

[OECD 443]. The omission of a second generation was substantiated by an extensive evaluation performed by two research groups [Janer et al. 2007; Martin et al. 2009; Piersma et al. 2011; Rorije et al. 2011] that concluded that a second generation will not change the risk assessment or classification and labeling significantly. This was reported as OECD document no. 176 [OECD No. 176, 2012].

Still, there are some discrepancies in the study design that is requested by the different regulatory authorities. On one hand, the US and Canada are following the OECD 443 design (including the developmental neuro-and immunotoxicity cohorts by default) and follow OECD guidance no. 117 for the use of internal triggers for inclusion of the second generation [OECD GD 117, 2011]. On the other hand, ECHA in Europe is requesting the EOGRTS without developmental neuro-and immunotoxicity cohorts as default (but these cohorts could be added based on available data) and no internal triggers for inclusion of the second generation are allowed. ECHA only agrees on a second generation beforehand (based on available data) or in case of equivocal findings during the first generation [ECHA R.7a, 2016].

The determination of triggers for a second generation as presented in OECD guidance no. 117 contradicts with the previously mentioned convincing publications on the relevance of the second generation. The reason why the US and Canada are using internal triggers for inclusion of the second generation is because they are still cautious and concerned that important information would be missed in case the second generation is omitted by default. OECD guidance no. 117 mentions "*Under the Canadian Pest Management Regulatory Agency and the United States Environmental Protection Agency, Office of Pesticide Programs, the decision to assess a second generation will take into consideration all available information including data gathered during the conduct of the EOGRTS. It is expected that, with the additional parameters evaluated in the F_1 generation in the EOGRTS, the F_2 with their limited parameter assessments would seldom affect the hazard characterization for risk assessment. However, until sufficient empirical data can be provided to support this hypothesis, the decision by the Canadian Pest Management Regulatory Agency and the United States Environmental Protection Agency, Office of Pesticide Programs to assess the second generation will be determined through the implementation of triggers, as defined in this Guidance Document*" [OECD GD 117, 2011].

Importantly, OECD guidance no. 117 states "*When a maximum of 20 EOGRTS have been conducted with an assessment of the second generation, the available results and appropriate scientific evidence will be analyzed. Subsequently, the guidance document will be reviewed and, if necessary, revised in light of experience gained*" [OECD GD 117, 2011]. Overall, one might appreciate that the addition of 20 EOGRTS would add just a small part to the evaluation of the hundreds already available multi-generation studies.

Of note, there is also concern that a second generation mating based on first generation triggers has the potential to miss signals of developmental or reproductive toxicity, especially if a

test substance produces transgenerational effects that accumulate with time, such as has been hypothesized with certain epigenetic changes and endocrine disruptors. The effects on the first generation may not be of sufficient magnitude to trigger an evaluation of a second generation. For this reason, some assert that a second generation mating should be made automatic rather than a triggered event. Certainly, if previous data for a test substance are equivocal or suggest potential transgenerational effects, then it is recommended that a second generation mating be included as part of the EOGRTS design.

Out of the 3R's, replacement is the most difficult one. Nonetheless, a lot of dedicated work is being executed by many research groups worldwide. Existing alternative test methods in reproductive toxicology, although not implemented, can be divided into three categories with increasing complexity: established cell lines, primary cell cultures, and mammalian embryos [Piersma, 2006].

1) Continuous cell lines have the exclusive advantage that they are devoid of the use of experimental animals (e.g. the mouse embryoic stem cell test, mEST). At the same time, their major disadvantage is their relative simplicity as they allow the study of effects on single mechanisms only, e.g. cell proliferation or cellular differentiation. In addition, these cell lines can loose their intrinsic capacities over multiple usages.
2) Primary cell cultures (e.g. limb bud or neural crest cells) require animal material, but they have the advantage that these cells have not been immortalized which may have possible adverse consequences for responses in test systems.
3) Whole embryo culture (WEC) systems offer the most complete in vitro alternatives for animal testing. They do require animal material and are more laborious but have the advantage of representing embryogenesis in its full complexity from cellular proliferation and differentiation to pattern formation.
4) Human cell lines/cultures are more human relevant and are therefore preferred for human risk evaluation. During the last decade, the application of human embryonic stem cells (hESC) in toxicity testing has been extensively studied and differentiation assays have been developed [Adler, 2008]. However, these methods are still under development.

Over 15 years ago, a full validation study of three in vitro assays (i.e. the rat limb bud micromass assay, rat WEC, and the mEST was conducted in four laboratories [Genschow et al. 2002]. A total of twenty chemicals and a positive control were tested; chemicals were correctly categorized in 78% of experiments using the mEST, 80% in WEC, and 70% in the rat limb bud micromass assay. However, these have never been implemented by regulatory agencies, and the role of in vitro assays as screens or as replacements for animal testing is still being determined. Moreover, no further in vitro DART assays have been validated since. Currently, these assays are used to screen compounds for further development as drug candidates or to prioritize chemicals for animal testing; they may also be used for mechanistic studies. Nevertheless, some believe that these tests should replace animals, particularly the second species required for evaluation of embryo/fetal

development [Lee, 2012]. Others believe that acceptance of one-in-one replacement of complex in vivo tests by relatively simple in vitro assays is not feasible. Rather, a mechanistically based alternative approach using the Adverse Outcome Pathway concept could stimulate further (regulatory) acceptance of non-animal tests [Piersma, 2014].

In addition to the in vitro assays described above, the zebrafish embryotoxicity test (ZET) has gained more consideration as it provides one major advantage when compared to the other assays, i.e. the ZET enables continuous developmental monitoring of a whole organism from early embryogenesis until hatching and beyond.

The zebrafish (ZF) model is not regulated by current legislations on animal welfare up to 120 hours post-fertilization (hpf). This is because a license is required on fish from the time at which they become capable of independent feeding, which for the majority of ZF is at 5 days post-fertilization (dpf) [Fleming 2007]. Therefore, the use of ZF embryos is considered a replacement of current animal experiments for toxicity and teratogenicity. Moreover, where present models for teratogenicity, usually rats and rabbits, require sacrifice of the dam to obtain the embryos for evaluation of developmental toxicity at a single time point, the ZF model allows continuous developmental assessment without sacrifice of the maternal fish. ZF organogenesis is completed within 72 hpf and its development is similar to mammalian embryos [Westerfield 1995]. Moreover, the embryo's transparency throughout organogenesis into the larva stage, facilitates detailed evaluation of structures and organ systems in an intact non-mammalian vertebrate embryo at all time-points during its complete embryogenesis. Not only the rapid organogenesis of the transparent ZF embryo, but also the simplicity of maintaining and breeding the adult ZF, makes this animal model ideal for developmental toxicity screening.

The predictability of the ZF model is highly influenced by compound uptake and thus internal exposure. Therefore, the mentioned advantages are counterbalanced by uncertainties about internal exposure and related sensitivity of the system. It is thought that the chorion is an important barrier for penetration of compounds until hatching. This means that several physicochemical properties of compounds tested with ZF embryos are important for generation of valid results. Important factors are size, molecular weight, lipophilicity, and the presence or absence of active transport into and within the embryo. The chorion of ZF embryos contains pores that have an average size of 0.17 μm^2 [Rawson 2000, Cheng 2007, Lee 2007], which restricts uptake of compounds exceeding the size of 3 kDa or 3000 g/mole for embryos younger than 24 h and might at least partiallyreach 4 kDa in 48 h embryos [Creton 2004; Pelka 2017]. Moreover, the uptake of compound through the chorion and into the embryo is thought to be enhanced by use of solvent within the test solution. Several studies found an effect of dimethyl sulfoxide (DMSO) on compound uptake and speculated that it might affect the predictability of the test [Hallare 2006, David 2012, Kais 2013]. One of the most important factors for internal exposure is thought to be lipophilicity expressed as logP value, of a compound. The logP has already been linked to bioavailability of chemicals in adult fish [Fu 2009, Schreiber 2009, Arnot 2010, Peterson 2011]. In addition, chemicals with higher logP, and

thus more lipophilic, tended to be more often toxic to the developing ZF than chemicals with lower logP [Padilla 2012]. In addition, only a few chemicals with logP below 0 were found to be toxic in ZF embryos [Steinberg 2013].

Since 2009, a consortium of six biopharmaceutical companies are collaborating to develop a harmonized ZET [Brannen 2010, Panzica-Kelly 2010, Gustafson 2012], but technical differences still remain and are the primary contributor to inter-laboratory differences in predictivity [Ball 2014]. So far, no global harmonized experimental setup has been established.

In 2014, the ZET was proposed as update in the ICH S5 guideline for preclinical testing of pharmaceuticals [ICH S5(R3), 2014]. For some time now, discussions are ongoing on the added value of a second species in developmental toxicity assessment (i.e., the use of both rats and rabbits in these preclinical tests). Future perspectives are aiming for a shift from the current two-species testing paradigm to an integrated testing strategy consisting of a combination of one or more in vitro or non-mammalian in vivo embryotoxicity tests and an in vivo embryo-fetal development study in one species. The ICH draft concept paper mentions that "a combination of in vitro assays, in silico prediction, and ZF assays is advocated rather than relying on the use of one stand-alone method." However, currently the most optimal study design has not been decided on and many different methods are being used by many research groups. This first needs to be harmonized before any possibility of implementation for regulatory purposes.

1.7. OBJECTIVE AND OUTLINE OF THIS THESIS

1.7.1. Objective of the thesis

Being based within a Contract Research Organization (CRO), I focused this thesis on the field work; aiming at practical implementations with the goal to instruct and motivate peers to reach a feasible 3R (reduce, refine and replace) standard within the currently accepted and performed study models for regulatory developmental and reproductive toxicity (DART) testing. Therefore I set my objective to assess whether and how improvements on animal welfare in DART testing can be made by refinement, reduction and/or replacement of regulatory requested studies, without loss of integrity and validity of the study designs.

1.7.2. OUTLINE OF THE THESIS

The work is divided into three parts:

The first part (Chapters 2-3) describes the implementation of refinement and reduction in current DART studies. This concerns refined blood sampling techniques which lead to less discomfort and reduced animal use. In **Chapter 2** we evaluated the use of capillary microsampling (CMS) in juvenile rats in support of reduced animal usage. The goal was to study the best location for CMS in

the juvenile rat (on postnatal day (PND) 4, 10 and 17) with the least discomfort to the animal and least impact on the study outcome. Utilizing the optimal method could decrease the number of satellite juvenile animals by 75%.

CMS can also be used in embryo-fetal development (EFD) studies in the rat, and would eliminate the need of satellite animals. However, in case CMS cannot be used, another method to omit satellite animals in EFD studies is the use of jugular vein blood sampling. This was successfully demonstrated in **Chapter 3**.

The second part (Chapters 4-7) depicts the impact on refinement and reduction in recently updated DART guidelines. The efficacy of reproduction/developmental screening studies (OECD 421 and 422 adopted in 1995 and 1996) was assessed in a retrospective evaluation of 134 studies (**Chapter 4**). It was concluded that reproduction/developmental screening studies are effective in providing unique data, especially considering the limited number of animals used. However, some simple additions would enrich their value in risk assessment even further. Some of the suggestions published in this paper were incorporated in the OECD guideline update through addition of endocrine relevant endpoints. Details on these additions are discussed in **Chapter 5**. This chapter provides important aspects of these challenging study designs, and was published with the aim to share our knowledge on their implementation in our laboratory so it can be used as guidance for other laboratories, study monitors and registration officers. The extra assessments were critically evaluated on their added value, with a special focus on the 3Rs.

Chapters 6 and 7 describe details on our experiences with implementation of OECD 443, the extended one-generation reproductive toxicity study (EOGRTS). The hallmark of the EOGRTS is that, based on certain criteria or triggers, selected offspring are assigned at weaning to different cohorts for further investigation of sexual maturation, reproductive organ integrity and function, neuropathological and behavioral endpoints, and/or immune function. The triggers allow for a more customizable design based directly on the data, while minimizing animal usage. Both chapters discuss the important aspects of this challenging study design. Chapter 6 focuses on practical implementation, while Chapter 7 offers guidance on science-based triggers for the extended evaluations (i.e. developmental neurotoxicity, developmental immunotoxicity and second generation mating).

The third part (Chapters 8-9) provides a first step towards replacement in future DART testing through optimization of the zebrafish embryotoxicity test (ZET). Currently many different methodologies exist for the performance of the ZET. To assist in harmonization, **Chapter 8** proposes the most optimal study design based on literature and empirical data, regarding temperature, test chamber, exposure period, presence of chorion, solvent use, exposure method, choice of concentrations, and teratogenic classification. Furthermore, our morphology scoring system was reported in detail to further enhance study design harmonization. In addition, **Chapter 9** describes the important role of internal zebrafish exposure on the predictability of the ZET. The impact of lipophilicity, solvent use and chorion presence on compound distribution within the zebrafish embryo were investigated.

Chapter 2

Effects of capillary microsampling on toxicological endpoints in juvenile rats

Xiaoyu Niu, Manon Beekhuijzen, Willem Schoonen, Harry Emmen, Mira Wenker

WIL Research Europe B.V., Hambakenwetering 7, 5231 DD 's-Hertogenbosch, The Netherlands.

Published in: Toxicological Sciences, 2016, Nov;154(1):69-77. Epub 2016 Aug 10. doi: 10.1093/toxsci/kfw146

ABSTRACT

Blood sampling during juvenile rat toxicology studies is required to determine the toxicokinetic (TK) profile of compounds. Juvenile rats are too small to undergo repeated blood sampling using conventional methods, which collect 200-300 µL blood at each time point. Recently, capillary microsampling (CMS) gained interest since sample sizes are almost 10 times smaller enabling multi-sample collection from one rat. Here, we evaluated the use of CMS in juvenile rats in support of reduced animal usage.

Juvenile rats at postnatal day (PND) 4, 10 and 17 underwent CMS via the submandibular, tail and jugular veins. The CMS methods for pups at different ages were evaluated based on sample quality and technical practicality as well as on acute and chronic changes of toxicological parameters.

The best location for CMS was the submandibular vein for PND 4 and 10 pups and the tail vein for PND 17 pups. No effects were found on clinical signs, body and organ weights and biochemistry parameters when 2x32 µL of blood was withdrawn from PND 4 pups or when 3x32 µL was taken from PND 10 and 17 pups within 24 hours. Significant changes in several hematology parameters were observed 24 hours after CMS due to a decrease of red blood cells and renewed production. These values had recovered to normal 7 days after CMS. CMS is feasible in juvenile rats for TK assessment. Utilizing this method could decrease the number of additional animals by 75%.

Keywords
Capillary microsampling (CMS), 3Rs (reduce replace refine), preclinical, toxicology, toxicokinetics

2.1. INTRODUCTION

Non-clinical toxicity studies in juvenile rats are required by the FDA [Food and Drug Administration, 2013] and EMA [European Medicines Agency, 2013] in advance of clinical trials in pediatric populations. For these (non-clinical) toxicity studies the assessment of the toxicokinetic (TK) behavior of the tested compound is essential, in order to correlate circulating drug concentrations with functional or pathological changes in animals during the dosing period. For this TK assessment, the traditional blood volume withdrawn at each time point is 200-300 µL. Due to the small size of juvenile rats (at postnatal days 1-21 the body weight ranges from 6 - 45 g), a terminal procedure is used to collect a single or half blood sample. Therefore, additional animals are included in juvenile toxicity studies for TK blood collection. Moreover, a full TK profile requires repeated blood sampling, which means the number of additional animals can often exceed the number of main test animals. In general, the collection of blood samples using additional groups in rat studies has been identified as the largest determinant of rodent numbers used in non-clinical safety studies [Sparrow *et al.*, 2011].

Capillary microsampling (CMS) is a relatively new technique for obtaining small amounts of blood (~ 30 µL) using an EDTA coated capillary. It was developed in 2012 at AstraZeneca [Jonsson *et al.*, 2012a]. Since then, it has been validated by several organizations, pharmaceutical companies and contract research organizations [Chapman *et al.*, 2014; Nillson *et al.*, 2013]. It has already been applied to obtain repeated blood samples from adult rats, mice and dogs [Dillen *et al.*, 2014; Jonsson *et al.*, 2012a,b; Korfmacher *et al.*, 2015; Luo *et al.*, 2015; Powles-Glover *et al.*, 2014a; Sadler and Bailey, 2013]. These data have paved the way for adopting CMS in main test animals, thereby reducing or even preventing the use of additional animals for TK blood collection, and allowing a direct link of exposure data to clinical and pathological observations for individual animals [Chapman *et al.*, 2014]. Applying CMS for juvenile toxicity studies has the potential to reduce the number of animals required per study even more dramatically than in adult animals. However, it is technically more challenging.

The only published data using CMS in juvenile rats is on postnatal day (PND) 19 pups [Powles-Glover *et al.*, 2014b] and on PND 9, 14 and 21 [Clubb and Jardine, 2015]. In the current study, we applied CMS on pups younger than 18 days. Since PND 4 is in general the youngest age for dosing rats, CMS first was performed on PND 4 pups. To gain experience with a wide range of ages, the youngest age (PND 4) plus one week (PND 10) and two weeks (PND 17) were also investigated. Furthermore, we studied the best route for blood collection (retro-orbital, saphenous, submandibular, tail and/or jugular vein) and whether CMS can be performed once, twice or three times within 24 hours. In addition, the quality of the blood samples was assessed with respect to the amount of blood collected and the occurrence of hemolysis and blood clotting. Acute (24 h) and chronic (7 days) effects of CMS on clinical signs, body and organ weights, hematology and biochemistry parameters indicative for toxicity were analyzed.

Overall, we evaluated the utility of CMS in the conduct of juvenile toxicology studies and the consequential reduction in animal use.

2.2. MATERIAL AND METHODS

2.2.1. Animals and group size

To obtain the pups for the present study, time-mated female Wistar Han Rats (Crl:WI(Han) (outbred, SPF-Quality)) from Charles River Deutschland (Sulzfeld, Germany) were used. On PND3, each litter was rearranged to 5 pups per gender. The main study consisted of three groups. Ten litters (40 males and 40 female pups) were allocated to the PND4 study. Ten litters (20 males and 20 female pups) were allocated to the PND10 study. Ten litters (40 males and 40 female pups) were allocated to the PND17 study. Experimental groups PND 4, 10 and 17 were all divided into two subgroups, A (Acute) and C (Chronic). Animals from subgroup A were used to evaluate acute effects: all pups were sacrificed 24 hours after the first CMS time. Animals from subgroup C were used for chronic effect evaluation, where all pups were sacrificed 7 days after the first blood withdrawal. Concurrent control groups where no sampling was performed were included for each of the ages in the study design. The study design and group PND4, 10 and 17 division and segregation is further described in Table 2.1. A power analysis to calculate the sample size based on the known information of red blood cells in suckling rats [Powles-Glover *et al*., 2014b] was performed. Based on this, a group size of 10 (5 females and 5 males) was chosen.

The experiments were approved by the Animal Welfare Officer and the Ethical Committee at WIL Research Europe B.V., as required by the Dutch Act on Animal Experimentation (February 1997).

2.2.2. Housing and husbandry

Animals were kept under standard conditions. Pups were kept with their dams in Macrolon plastic cages (MIII type, height 18 cm) until termination (except for pups from the PND 17C group which were kept with their dams until PND 21, and were group housed (5/sex/group) until termination on PND24). Environmental controls for the animal room were set to maintain 18 to 24°C, a relative humidity of 40 to 70%, at least 10 air changes/hour, and a 12-hour light/dark cycle. Animals had free access to pelleted rodent diet (SM R/M-Z from SSNIFF® Spezialdiäten GmbH, Soest, Germany) and tap water. The animals were further supplied with sterilized sawdust as bedding material (Lignocel S 8-15, JRS - J.Rettenmaier & Söhne GmbH + CO. KG, Rosenberg, Germany) and paper as cage-enrichment and nesting material (Enviro-dri, Wm. Lillico & Son (Wonham Mill Ltd), Surrey, United Kingdom).

2.2.3. CMS in juvenile rats at PND4, 10 and 17

Four CMS methods were tested: the saphenous vein, submandibular vein (SV), lateral tail vein (TV), or jugular vein (JV). A pilot study was performed to select the suitable CMS methods for animals at three different ages (PND4, 10 and 17). However, saphenous vein sampling was not performed further due to poor visibility of the vein in all pups. The study design for the main study animals presented in Table 2.1 was based on the results obtained from the pilot study.

Table 2.1. Scheme of test set-up.

Animals	Groups		Numbers		Micro-sampling frequency	Time points (hours post dose[a])			Termination (PND)	Microsampling techniques		
			M	F	32 µL	0.5	2	4		SV	TV	JV
PND 4	**A**	Control	5	5	-	-	-	-	5	-	-	-
		Acute	5	5	1	S	-	-	5	S	-	-
	C	Control	5	5	2	S	S	-	5	S	-	-
		Chronic	5	5	3	S	S	S	5	S	-	-
			5	5	-	-	-	-	11	-	-	-
			5	5	1	S	-	-	11	S	-	-
			5	5	2	S	S	-	11	S	-	-
			5	5	3	S	S	S	11	S	-	-
PND 10	**A**	Control	5	5	-	-	-	-	11	-	-	-
		Acute	5	5	3	S	S	S	11	S	-	-
	C	Control	5	5	-	-	-	-	17	-	-	-
		Chronic	5	5	3	S	S	S	17	S	-	-
PND 17	**A**	Control	5	5	-	-	-	-	18	-	-	-
		Acute	15	15	3	S	S	S	18	S	S	S
	C	Control	5	5	-	-	-	-	24	-	-	-
		Chronic	15	15	3	S	S	S	24	S	S	S

M = Male; F = Female; SV = submandibular vein; TV = tail vein; JV = jugular vein; PND = Post Natal Day; A = Acute; C = Chronic; - not applicable, not sampled; S = sampled.

[a]Animals were not actually dosed and so "time after dose" refers to time after 8 am, which would be a typical time of dosing in a once daily dosing study.

From a study design perspective, optimizing blood-sampling regimes is technically demanding in young animals. Too much blood loss could potentially compromise animal welfare. According to the NIH-ARAC guideline on survival bleeding, rodents can recover in 7 days when less than 7.5% CBV (circulating blood volume) is taken. With this criterium, the calculated safe blood withdrawal volume for PND 4, 10 and 17 pups is 33-58 µL, 82-130 µL and 140-210 µL, respectively. It was reported that taking a 6 × 32 µL CMS from pups on PND 19 in the 24 h preceding necropsy was possible [Powles-Glover *et al.*, 2014a]. We limited blood collection to 3 times (3x32 µL) as the highest frequency for the PND 10 and 17 pups. Blood sampling (1, 2, or 3 times) was tested in PND 4 pups. Microsampling was carried out according to a toxicokinetic blood sampling regime. Blood (32 µL) was collected using capillaries (K2-EDTA treated hematocrit tubes, Vitrex® Micro hematocrit tubes) at each time point. The capillary was plugged at one end with wax (Vitrex Sigillum Wax, Vitrex medical A/S, Denmark) and placed in an appropriately labeled tube for further use.

PND4 animals were microsampled via the submandibular vein. Animals were restrained manually. The submandibular vein was punctured with a 25G needle (Terumo Europe N.V., Belgium). When the blood started to flow, the capillary was kept at the vein puncture. The capillary was filled until the first mark (approximately 32 µL). Absorbable Haemostatic Gelatin Sponge (SPONGOSTAN™, Ferrosan Medical Devices A/S, Denmark) was applied to the bleeding surface for hemostasis. Blood was sampled 1 to 3 times in 24 hours following a toxicokinetic blood sampling regime described in Table 2.1. Repeated sampling was possible by alternating the side of the cheek. PND4 pups could not be microsampled via tail or jugular vein, since these veins were not visible.

PND10 animals were microsampled via the submandibular vein. The restraining, needle punch, blood collection and hemostatic methods were similar to those for the PND4 pups. However, a 21G needle was used for PND10 animals. Blood was sampled 3 times in 24 hours at 0.5, 2 and 4 hour time points. PND10 pups could not be microsampled via tail or jugular vein, since the jugular vein was not visible and the tail vein did not yield enough blood for CMS collection.

PND17 animals were microsampled via the submandibular, tail or jugular vein. Animals were transferred to an incubator set at 39 °C for approximately eight minutes prior to blood sampling from the tail vein. Pre-warming was not necessary for the other two CMS methods. Animals were restrained during blood sampling by tubes (tail vein) or manually (submandibular vein and jugular vein). The submandibular vein was punctured with a 23G needle, when the blood started to flow, a glass capillary was directly brought into contact with the blood at the vein puncture. A 23G needle in direct contact with a capillary container was used for the jugular vein CMS. The capillary was kept inside the container at the end of the needle, the needle tip was held inside of the vein until enough blood was collected into the capillary container. A butterfly needle (Terumo SurFlo Safety Winged Infusion Sets 25 Gauge) was inserted into the tail vein, the capillary was connected directly to the silastic tube end of the butterfly needle. Gentle finger (or tissue) pressure over the puncture site stopped the bleeding in a few seconds for all three methods. Blood was sampled 3 times in 24 hours at 0.5, 2 and 4 hour time points.

2.2.4. Clinical observations and in-life monitoring

All pups were thoroughly examined after birth. Clinical signs were observed at least once daily. Mortality and viability were checked at least twice daily. Body weights were monitored for group PND 4C, 10C and 17C animals from the first microsampling day until one day before the necropsy day.

2.2.5. Hematology and clinical biochemistry measurement

Blood samples drawn at termination were collected into tubes (Greiner Bio-One GmbH, Kremsmünster, Austria) prepared with K_3-EDTA for hematological parameters (0.5 mL), and/or tubes treated with Li-heparin (Greiner Bio-One GmbH, Kremsmünster, Austria) for clinical biochemistry parameters (0.5 mL).

The following hematology parameters were determined in blood samples using the ADVIA® 120 Hematology System (Siemens Healthcare Diagnostics B.V., Breda, The Netherlands): White blood cells (WBC), Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils, Red blood cells, Reticulocytes, Red blood cell distribution width (RDW), Hemoglobin, Hematocrit, Mean corpuscular volume (MCV), Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Platelets.

The following clinical biochemistry parameters were determined in plasma samples using the AU400 (Beckman Coulter Nederland B.V., Woerden, The Netherlands): Alanine aminotransferase (ALAT), Aspartate aminotransferase (ASAT), Alkaline Phosphatase (ALP), Total protein, Albumin, Total globulin, Albumin Globulin ratio, Total Bilirubin, Urea, Creatinine, Glucose, Cholesterol, Triglycerides, Phospholipids, Sodium, Potassium, Chloride, Calcium, Inorganic Phosphate (Inorg. Phos).

2.2.6. Hemolysis measurements

The capillaries containing the microsampled blood were stored in a secondary labeled container suitable for centrifugation. Within maximally 2 hours after collection, the capillaries with blood were centrifuged at 2490g for 15 minutes at 4°C. After centrifugation, the color of the plasma was checked visually and the hemolysis level of each sample was recorded (scale: N = negative, + = slightly hemolytic, ++ = moderately hemolytic, +++ = severely hemolytic).

2.2.7. Termination procedures

PND5 pups were sacrificed by decapitation and whole blood was collected. The rest of the pups were deeply anaesthetized using isoflurane (Abbott B.V., Hoofddorp, The Netherlands) and blood was taken from the aorta from each animal for hematology and clinical biochemistry examination; subsequently these animals were exsanguinated.

Tissue samples from spleen and liver were taken from PND 4C, 10C and 17C animals. All retained organs were weighed, preserved in 10% buffered formalin (neutral phosphate buffered 4% formaldehyde solution, Klinipath, Duiven, The Netherlands). All spleens were examined histologically.

2.2.8. Statistical analysis

The body weights, organ weights, hematological and clinical biochemistry data were analyzed using the Dunnett-test when the data were normally distributed and Steel-test when the data were not normally distributed. All analyses were performed separately for males and females. The microsampled groups were compared independently with their concurrent age matched controls. All statistical tests were two-sided and were performed at the 5% and 1% significance level.

2.3. RESULTS

2.3.1. CMS methods selection during pilot study

In the pilot study, four CMS methods (saphenous, submandibular, tail and jugular veins) were tested according to guidelines for survival bleeding of mice and rats. Saphenous veins were invisible in all pups. The final three CMS methods were selected for rats on PND4, 10 and 17. Relevant parameters such as pre-warming, animal restraining, needle size, needle punch/insert, and hemostatic method were optimized as well. The summary of the method selection is shown in Table 2.2.

2.3.2. Quantity and Quality of the blood samples by CMS

Sufficient blood samples could be drawn from all animals at each time point by all selected CMS methods. Repeated CMS did not influence the amount of blood withdrawal. None of the blood samples showed hemolysis.

2.3.3. Clinical observations

In the PND 17 submandibular vein group, 3 out of 20 pups displayed clonic spasms during blood sampling (one male pup displayed clonic spasms after one time sampling, one female after two times sampling, one male after three times sampling). In PND 4 submandibular vein group, all pups showed bruises on the cheek after one time sampling. In 12 out of 30 pups the hematomas continued until the end of the sampling day. The hematomas disappeared the day after in all pups. No procedure related clinical observations were recorded in the rest of the experiments.

2.3.4. Body weights

Body weights of animals from the PND 4C, 10C and 17C groups are presented in Figure 2.1. Body weights of microsampled animals remained in the same range as controls over the recorded period for both males and females in all the groups.

2.3.5. Hematology

The hematology parameters were compared between microsampled pups and their concurrent age-matched controls. The statistically significant changes are summarized in Table 2.3. Of all the collected data, the only changes found were a reduction of hemoglobin, hemotocrit and red blood cell numbers in the treated group of PND 10A and 17A, while an increase in reticulocytes was seen in treated groups of PND 10A. In the PND 10C and 17C group only slight effects on MCHC levels were found.

The terminal blood sample was less than 0.5 mL from each PND 5 pup. The hematology data were not available for the PND 4A group since the majority of the blood samples (30/40) collected at necropsy were clotted. Some of the clot free blood samples did not contain enough volume for a proper hematology measurement (less than 300µL); results from these were not reported.

Table 2.2. Summary of CMS methods selection during pilot study

	PND 4	PND 10	PND 17		
Method	Submandibular vein	Submandibular vein	Submandibular vein	Tail vein	Jugular vein
Prewarming	No	No	No	39°C incubator 8 min	No
Shaving	No	No	No	No	Yes
Restraining Needle size	Hand restrain 25G punch	Hand restrain 21G punch	Hand restrain 23G punch	Tube restrain 25G butterfly insert[b]	Hand restrain 23G insert[b]
Hemostasis	Hemostatic gelatin sponge	Hemostatic gelatin sponge	Finger/tissue Compression	Finger/tissue Compression	Finger/tissue Compression
Frequency	3 x 32µL[a]	3 x 32µL[a]	3 x 32µL[a]	3 x 32µL	3 x 32µL
Time consumed	2.6 min/sample	1.6 min/sample	1.3 min/sample	2.3 min/sample	1.8 min/sample
Technicans (n)	2	2	2	2	2

[a]Repeated sampling (3 x 32µL) was possible by alternating the side of cheek
[b]The needle was inserted into the vein, and the capillary was kept at the end of the needle

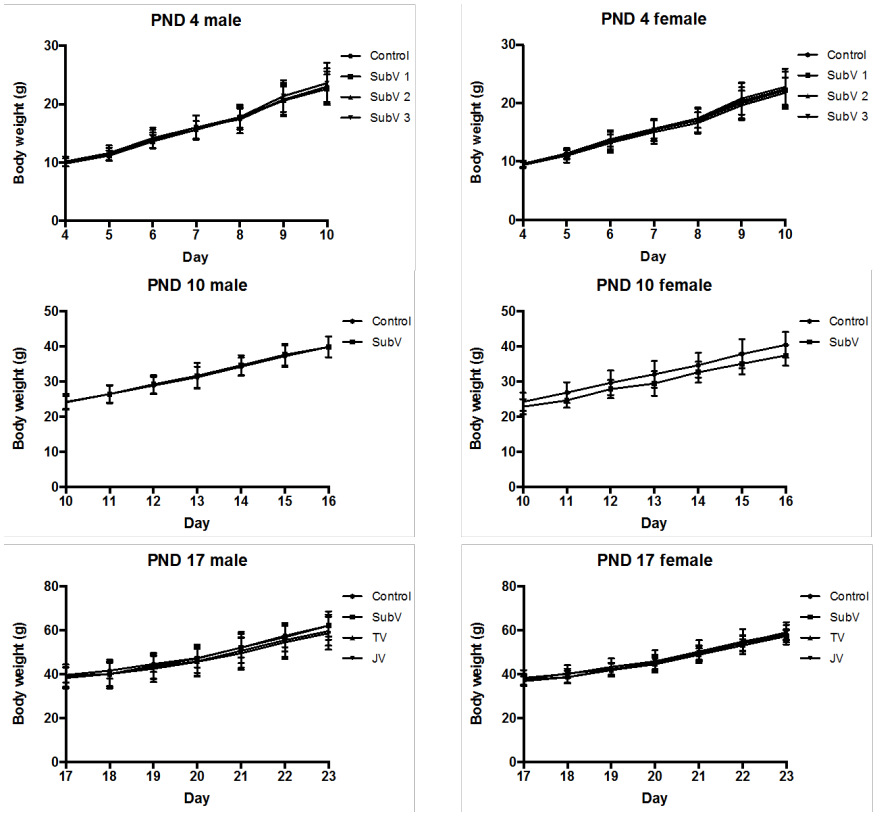


Figure 2.1. Bodyweights of microsampled and control animals from the PND4C, PND10C and PND17C groups. Data were recorded from the CMS day until one day before necropsy. The graph shows group mean (± SD) bodyweights for male and female animals (n=5). Data were compared between the CMS groups and their concurrent controls. No statistical significant differences were found for any of the groups.

2.3.6. Plasma biochemistry parameters

Plasma biochemistry data were compared between microsampled pups and the controls. The statistically significant changes were summarized in Table 2.4. Only some effects were observed in PND 10A, C and 17A, C treated groups with an increase or decrease of triglycerides, an effect on glucose and ALAT. The latter findings were only found in males. The biochemistry parameters were not measured for the PND 4A group because too little (150-300 µL) blood was available from each pup at termination.

2.3.7. Selected organ weights and histopathology

Liver and spleen weight were measured on the termination day in the chronic effect groups: PND 4C, 10C and 17C. There were no procedure-related organ weight changes associated with the blood sampling in both males and females compared to controls. In addition, no macroscopic abnormalities or hematopoietic score differences were observed in any of the pups in the livers and spleens. Absolute organ weight data and hematopoietic scores are summarized in Table 2.5.

Table 2.3. Summary of hematology outcome

Group	Termination	Gender	Changed Hematology Parameters	Control Mean ± SD	Treatment Mean ± SD	Changes %
PND 4A	5	M	-	-	-	-
		F	-	-	-	-
PND 4C	11	M	↓WBC (10E9/L) (SV3) NF (SV2 & SV1)	3.10 ± 0.40	2.20 ± 0.30*	29.0
		F	↓ Hemoglobin (mmol/L) (SV3) NF (SV2 & SV1)	5.70 ± 0.20	5.10 ± 0.40*	10.5
PND 10A	11	M	↓ Red blood cells (10E12/L) (SV) ↓ Hemoglobin (mmol/L) (SV) ↓ Hematocrit (L/L)(SV) ↓ Monocytes (%WBC) (SV) ↑ Reticulocytes (%RBC) (SV)	3.39 ± 0.24 5.80 ± 0.40 0.29 ± 0.02 6.40 ± 2.70 26.40 ± 1.80	2.65 ± 0.23** 4.60 ± 0.30** 0.23 ± 0.02** 3.60 ± 1.30^{++} 31.20 ± 2.00^{+}	21.8 20.7 20.2 43.7 18.2
		F	↓ Red blood cells (10E12/L) (SV) ↓ Hemoglobin (mmol/L) (SV) ↓ Hematocrit (L/L)(SV) ↓ Monocytes (%WBC) (SV) ↑ Reticulocytes (%RBC) (SV)	3.32 ± 0.19 5.70 ± 0.30 0.30 ± 0.02 7.90 ± 2.20 26.60 ± 2.10	2.73 ± 0.30** 4.60 ± 0.40** 0.24 ± 0.02** 3.30 ± 2.10^{++} 32.00 ± 2.30^{+}	17.8 19.3 18.9 58.2 20.3
PND 10C	17	M	↓ MCHC (mmol/L) (SV)	20.77 ± 0.23	20.09 ± 0.19**	3.3
		F	↓ MCHC (mmol/L) (SV)	20.54 ± 0.29	19.65 ± 0.79*	4.3
PND 17A	18	M	↓ Red blood cells (10E12/L) (TV) ↓ Red blood cells (10E12/L) (JV)	4.13 ± 0.21 4.13 ± 0.21	3.75 ± 0.14* 3.71 ± 0.12*	9.2 10.2
		F	↓ Neutrophils (%WBC) (SV) ↓ Red blood cells (10E12/L) (SV) ↓ Hemoglobin (mmol/L) (SV) ↓ Hematocrit (L/L) (SV) ↓ MCHC (mmol/L) (SV) ↓ MCHC (mmol/L) (JV) ↓ MCHC (mmol/L) (TV)	16.80 ± 4.50 4.26 ± 0.40 5.70 ± 0.50 0.27 ± 0.02 20.58 ± 0.33 20.58 ± 0.33 20.58 ± 0.33	10.20 ± 2.80^{+} 3.50 ± 0.20* 4.50 ± 0.30* 0.23 ± 0.02* 20.01 ± 0.20* 19.84 ± 0.20** 19.71 ±0.31**	39.3 17.8 21.1 17.5 2.8 3.6 4.2
PND 17C	24	M	↑ Monocytes (%WBC) (TV)	2.10 ± 0.40	2.60 ± 0.30^{+}	23.8
		F	NF	-	-	

Data were compared between the microsampled groups and their concurrent controls.
-: not applicable; NF: no findings. For details, see supplementary tables.
↑ increase ↓ decrease
+/++ Steel-test significant at 5% (+) or 1% (++) level.
*/** Dunnett-test based on pooled variance significant at 5% (*) or 1% (**) level.
SV: submandibular vein, JV: jugular vein, TV: tail vein, SV3: 3 times CMS via submandibular vein.

Table 2.4. Summary of biochemistry outcome

Group	Termination	Gender	Changed Biochemistry Parameters	Control Mean ± SD	Treatment Mean ± SD	Changes %
PND 4A	5	M	-	-	-	-
		F	-	-	-	-
PND 4C	11	M	NF	-	-	-
		F	NF	-	-	-
PND 10A	11	M	↑ Triglycerides (mmol/L)(SV)	3.20 ± 0.74	4.88 ± 0.46**	52.5
		F	NF	-	-	-
PND 10C	17	M	↑ Glucose (mmol/L)(SV)	9.61 (0.52)	10.35 ± 0.24*	7.7
		F	NF	-	-	-
PND 17A	18	M	NF	-	-	-
		F	↓Triglycerides (mmol/L)(SV)	3.37 ± 0.82	2.18 ± 0.33*	35.3
			↓Triglycerides (mmol/L)(TV)	3.37 ± 0.82	2.40 ± 0.35*	28.8
			↓Triglycerides (mmol/L)(JV)	3.37 ± 0.82	2.44 ± 0.50*	27.6
PND 17C	24	M	↑ALAT (U/L)(SV)	38.2 ± 4.6	49.2 ± 9.2*	28.8
		F	NF	-	-	-

Data were compared between the microsampled groups and their concurrent controls.
-: not applicable; NF: no findings. For details, see supplementary tables.
↑ increase ↓ decrease
*/** Dunnett-test based on pooled variance significant at 5% (*) or 1% (**) level.
SV: submandibular vein, JV: jugular vein, TV: tail vein.

2.4. DISCUSSION

For the proper selection of the best microsampling technique, the most frequently used blood sampling sites in rodents with good survival rates such as retro-orbital, submandibular, saphenous, tail, and jugular vein (http://oacu.od.nih.gov/ARAC/documents/Rodent_Bleeding.pdf) were considered. The age and size of the animals were taken into account to determine the amount of blood that can be drawn compared to body weight. Moreover the type and quality of the sample required (*e.g.* serum, whole cells, sterility, tissue fluid contamination, etc.), as well as the frequency of sampling need to be assessed before the start of the experiment. Finally, also the animal health and welfare and the training and experience of the technician are crucial in this aspect.

- For retro-orbital sampling general anesthesia is highly recommended. However, this technique is ethically and esthetically less acceptable, requires a highly skilled phlebotomist, and has a greater potential than other blood collection routes to result in complications. Therefore, retro-orbital bleeding was excluded from this project.

Table 2.5. Chronic effects of capillary microsampling on changes in liver and spleen weights and hematopoietic scores between control and treatment groups of PND 4, 10 and 17 of both male and female pups.

Animals		Males		Females	
		Weight changes (g)			
		Liver Mean ± SD (n)	Spleen Mean ± SD (n)	Liver Mean ± SD (n)	Spleen Mean ± SD (n)
PND 4C	Control	0.67 ± 0.13 (5)	0.153 ± 0.030 (5)	0.67 ± 0.12 (5)	0.131 ± 0.032 (5)
	SV1	0.68 ± 0.10 (5)	0.148 ± 0.043 (5)	0.66 ± 0.11 (5)	0.135 ± 0.020 (5)
	SV2	0.64 ± 0.09 (5)	0.145 ± 0.024 (5)	0.68 ± 0.12 (5)	0.134 ± 0.031 (5)
	SV3	0.66 ± 0.11 (5)	0.136 ± 0.035 (5)	0.66 ± 0.11 (5)	0.134 ± 0.025 (5)
PND 10C	Control	1.16 ± 0.07 (5)	0.178 ± 0.017 (5)	1.20 ± 0.11 (5)	0.181 ± 0.011 (5)
	SV	1.15 ± 0.13 (5)	0.161 ± 0.019 (5)	1.16 ± 0.16 (5)	0.179 ± 0.032 (5)
PND 17C	Control	2.54 ± 0.08 (5)	0.276 ± 0.036 (5)	2.28 ± 0.19 (5)	0.293 ± 0.047 (5)
	SV	2.60 ± 0.19 (4)	0.285 ± 0.056 (4)	2.24 ± 0.08 (5)	0.265 ± 0.006 (5)
	TV	2.39 ± 0.17 (5)	0.268 ± 0.035 (5)	2.42 ± 0.30 (5)	0.280 ± 0.016 (5)
	JV	2.52 ± 0.31 (5)	0.280 ± 0.029 (5)	2.38 ± 0.37 (5)	0.295 ± 0.043 (5)
		Hematopoietic scores (1 to 4)			
		Liver Mean ± SD (n)	Spleen Mean ± SD (n)	Liver Mean ± SD (n)	Spleen Mean ± SD (n)
PND 4C	Control	1.00 ± 0.00 (5)	3.60 ± 0.55 (5)	1.40 ± 0.55 (5)	4.00 ± 0.00 (5)
	SV1	1.20 ± 0.45 (5)	3.60 ± 0.55 (5)	1.60 ± 0.55 (5)	3.60 ± 0.55 (5)
	SV2	1.20 ± 0.45 (5)	3.60 ± 0.55 (5)	1.60 ± 0.55 (5)	3.40 ± 0.55 (5)
	SV3	1.20 ± 0.45 (5)	3.40 ± 0.55 (5)	1.20 ± 0.45 (5)	4.00 ± 0.00 (5)
PND 10C	Control	1.00 ± 0.00 (5)	2.40 ± 0.44 (5)	1.00 ± 0.00 (5)	1.60 ± 0.55 (5)
	SV	0.60 ± 0.55 (5)	1.80 ± 0.45 (5)	1.00 ± 0.00 (5)	2.40 ± 0.89 (5)
PND 17C	Control	1.00 ± 0.00 (5)	3.00 ± 0.00 (5)	0.80 ± 0.45 (5)	2.80 ± 0.45 (5)
	SV	1.00 ± 0.00 (4)	3.00 ± 0.00 (4)	0.60 ± 0.55 (5)	2.80 ± 0.45 (5)
	TV	0.80 ± 0.45 (5)	3.20 ± 0.45 (5)	0.80 ± 0.45 (5)	3.20 ± 0.45 (5)
	JV	0.60 ± 0.55 (5)	2.80 ± 0.84 (5)	0.80 ± 0.45 (5)	3.40 ± 0.55 (5)

SV1, 2, 3: 1, 2 and 3 times CMS via submandibular vein.
SV: submandibular vein, JV: jugular vein, TV: tail vein

- The CMS of the saphenous vein was tried, but not used due to difficulties in visualizing the veins in all pups.
- Submandibular vein CMS was successfully performed for the pups at all three ages (PND 4, 10 and 17), and it was the only possible CMS method for PND 4 and 10 pups. More hands-on training was required to reliably withdraw a reasonable quantity of blood from PND 17 pups since their cheeks are covered by hair. For the observation of incidental clonic spasms in 3 out of 20 animals from 60 collections points, no clear explanation can be given, as two collection points with the same animal did not result in these spasms. Special care must be taken to ensure adequate hemostasis.
- Jugular vein CMS could easily be assessed in PND 17 pups. Without influencing animal welfare, 3x32 µL of blood was collected in a 24 hour regime. Shaving was required to reveal the jugular vein, and effective manual restraint was necessary. Needle insertion into the vein

prevented contamination of the blood samples with skin tissue. Repeated sampling was possible, as no clot or scab was observed after any sampling. Compared to tail vein and submandibular vein CMS, the jugular vein CMS required a higher degree of competence of the technician during animal restraint to reveal the vein properly, and during needle insertion to avoid harm to the rats.

- Lateral tail vein CMS was also feasible and successful for PND 17 pups. Pre-warming of the pups was necessary for this method. It was reported that PND 19 pups required warming at 38°C for 2-5 min prior to tail vein CMS [Powles-Glover, 2014a]. Furthermore, the time required in a hot box depended on the age of the pups. In this project, a pilot study using 5 min pre-warming for PND 17 pups resulted in insufficient blood yields. Therefore, 8 min pre-warming prior to blood-sampling was adopted for all the main study animals and it guaranteed successful blood withdrawal. Although no unusual behavior was observed in any of the study animals, warming induced stress could not be excluded since the entire animal was heated [Sadler and Bailey, 2013]. Blood sample collection via the tail vein was commonly performed by nicking the vessel [Powles-Glover et al., 2014a]. However, the quality of blood samples was variable due to contamination with tissue products. Butterfly needle inserts were used in this project to minimize the contamination without slowing down the sampling procedure (2.3 min/sample/one technician). Repeated collections were possible as no clots or scabs were observed after sampling.

Thus, the most preferred CMS method for PND 4 and 10 pups is the submandibular vein. For PND 17 pups the preferred method is via the tail vein since three animals showed convulsions after submandibular vein CMS, and the jugular vein CMS required a high degree of competence of the technician to get enough blood and to avoid harm to the rats.

The influence of CMS on hematological parameters revealed that removing 96 µL of blood in 24 hours from PND 10A and 17A pups resulted in statistically significant changes in some of the hematology parameters. In the animals sacrificed 24 hours after the microsampling, reduced red blood cell and monocyte (%WBC) counts, decreased hemoglobin and hematocrit levels, and increased reticulocytes were found for both males and females in PND 10A pups. This pattern of change was also observed in the PND 17A pups, but to a lesser extent. Decreased red blood cell counts were also seen in PND 17A female and male pups. Decreased hemoglobin and hematocrit levels were only found in PND 17A females. However, the decreased monocyte and increased reticulocytes found in PND 10A pups was no longer significant by PND 17A groups. This indicated that PND 10A pups were more affected by CMS induced blood loss than PND 17A pups. This can be expected considering the weight differences and the higher percentage of blood that was withdrawn from PND 10A pups compared to that of PND 17A pups. On the other hand, in the animals sacrificed 7 days after the microsampling, only decreased MCHC levels and increased monocytes were found in the PND 10C and 17C groups, respectively. As changes were slight and no corroborative findings were noted, they were not considered to be biologically relevant. It could therefore be concluded that the acute effects of CMS seen on hematology parameters in PND 10A and 17A pups recovered within one week.

With respect to the PND 4 pups more CMS related problems occurred. PND 4A pups could be microsampled 1-3 times. However, whole blood collection via decapitation for hematology examination was unsuccessful in this project. Therefore, acute effects of CMS on PND4A pups could not be assessed. Hematology data obtained from the PND 4C pups, terminated 7 days after the microsampling indicated a good recovery of pups, which were microsampled on PND 4 for one and two times, since no significant changes in hematology parameters were observed. On the other hand, in the pups which were microsampled three times, decreased WBC counts in males and decreased hemoglobin levels in females were observed. This indicated that animals which were microsampled 3 times did not fully recover within 7 days. Therefore, a maximum of 2 times CMS in 24 hours is suggested for PND 4 pups.

The effects of CMS on general clinical and biochemical parameters were also examined. Rat pups demonstrated that collection of 3 × 32 µL of blood from suckling pups on PND 4, 10 and 17, in a 24 h period, did not alter body and organ weights and biochemical blood parameters. Clinical signs were monitored during the experiments, and pups remained in good condition until the day of necropsy. Three PND17 pups were terminated after the submandibular vein CMS because of convulsions. However, this clinical observation was likely caused by the deep needle punch rather than by the blood removal. A bright color of the microsampled blood was observed, indicating artery damage. Moreover, all PND4 pups showed bruises on the cheeks after a single submandibular vein CMS, which recovered in all animals the day after. In all PND groups, both for males and females, the body weights remained in the same range as controls over the recorded period. Liver and spleen weights at 7 days after the CMS were comparable between the microsampled and control animals. These findings point towards normal body growth and indicate that the normal growth function of the pups is not affected by the CMS. With respect to the biochemistry parameters between the control and CMS groups, these changes were inconsistent, between males and females, between ages and between acute and chronic groups. Furthermore, no evidence for a biological effect, such as liver weight gain or liver pathology was found, causing changes in biochemistry parameters, as triglycerides, glucose and ALAT. Therefore, the changes in biochemistry parameters were considered to be within background and were not the side effects of the blood withdrawal.

The application of CMS to juvenile rats can greatly contribute to the 3R (Replacement, Reduction and Refinement). In a traditional preclinical juvenile rat toxicity study, blood samples (200-300 µL) acquired for TK exposure data are collected from additional animals via a terminal procedure. Our data showed that with CMS only 2 repeated blood collections from one pup from PND 4 can be carried out. Three collections do interfere with the PND4 pups well-being. In the case of PND 4 pups, the total blood volume obtained at termination from one animal is around 150-300 µL. In a traditional study, two pups are sacrificed to collect one TK sample at each time point. Normally 4 dose levels and 6 time points are included in a preclinical rat toxicity study. Thereby 3 animals of both sexes are used. Thus, 288 additional PND 4 pups are sacrificed to collect the TK blood samples. Using CMS, maximally 2 samples can be collected from one pup in 24 hours, which would reduce the additional pup numbers to 72. Thus, successful implementation of CMS will help to

greatly reduce the additional animal use in juvenile toxicology studies by 75%. However, as an acute effect of CMS on hematology parameters were seen, it is not suggested to use the main study animals for TK blood sampling.

2.5. CONCLUSIONS

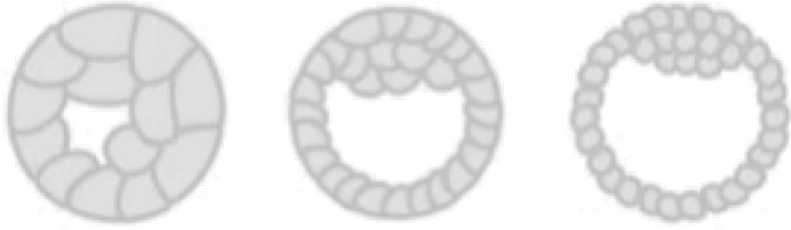
CMS is a more humane, efficient, and noninvasive method to collect blood samples. In traditional methods, a significant amount of blood coagulates inside the needle, resulting in poor yields. Because of the larger amount of blood, the wound needs to be larger and deeper to yield sufficient quantities of blood. This process is longer and animals may suffer more. CMS is relatively fast and no anesthesia is needed. Because the volume (32 µL) is small, the healing process is fast and does not interfere with the blood flow for the second sampling. No clot or scab needs to be removed when serial collection is required. Blood samples collected by CMS are of good quality; no hemolysis was observed in any of the microsampled blood samples. Also, sample quantity does not decrease with prolonged bleeding time and repeated CMS did not influence the amount of blood withdrawal.

The application of CMS on juvenile rats can greatly contribute to the 3R (Replacement, Reduction and Refinement). Successful implementation of CMS will help to greatly reduce the additional animal use in juvenile toxicology studies by 75%. However, as an acute effect of CMS on hematology parameters were seen, it is not suggested to use the main study animals for TK blood sampling. Nonetheless, CMS offers the opportunity for a large reduction in the number of animals used in juvenile toxicity studies.

Acknowledgements
We would like to thank the (bio)technicians, pathologists and photographer for their expertise and technical assistance.

Chapter 3

Satellite rats are redundant in embryo-fetal development studies



Manon Beekhuijzen, Yaomi de Lange, Ankie Lambregts, Birgit Peter, Mira Wenker, Harry Emmen

Charles River, 's-Hertogenbosch, The Netherlands.

Published in: Reproductive Toxicology. 2017 Sep;72:122-128. doi: 10.1016/j.reprotox.2017.06.042. Epub 2017 Jun 8.

ABSTRACT

Routinely in many laboratories, satellite rats are added to embryo-fetal development (EFD) studies for pharmaceuticals to assess toxicokinetic (TK) properties, because it is assumed that collection of multiple blood samples with relatively large volumes might affect the study outcome. With recent refinement of blood sampling techniques, this belief requires reevaluation. The current work showed successful implementation of jugular vein blood sampling in an EFD rat study without satellite animals, thereby reducing the number of rats in standard EFD studies for pharmaceuticals by 20%. Although not evaluated in this study, microsampling has shown to be very successful and eliminates the need of satellite animals. However, currently not all laboratories have implemented this method and regularly the bioanalytical method is already developed with a limit of quantification that is insufficiently sensitive. Therefore in those cases, a quick win to omit satellite animals can be established by using jugular vein blood sampling.

Keywords
Embryo-fetal development (EFD) study; 3R principles; jugular vein blood sampling; microsampling; toxicokinetic (TK); satellite animals

3.1. INTRODUCTION

Guidance document ICH M3 (R2) [ICH M3(R2), 2009] describes the type of nonclinical toxicity studies required for safety assessment of pharmaceuticals and its timing in respect of the clinical phases. Developmental toxicity is generally assessed in two species: a rodent and non-rodent (usually the rat and rabbit). In the EU and Japan definitive developmental toxicity studies in two species are required, in general, before inclusion of women of childbearing potential (WOCBP) in any clinical study. However, there are some exemptions to this. One circumstance could be a short-duration (e.g., 2 weeks) clinical study with precautions taken to prevent pregnancy. Another circumstance is where the clinical study must include WOCBP since the disease is prominent in women, but sufficient precautions are used to prevent pregnancy. In the relevant regions, the guidance also allows for the inclusion of WOCBP (up to 150) using adequate birth control methods for a relatively short duration (up to 3 months) if preliminary developmental toxicity data are available from two species. In the United States, assessments of embryo-fetal development (EFD) can be deferred up to but not including phase 3 clinical studies for WOCBP as long as precautions are taken to prevent pregnancy.

Details on the study design are described in ICH guideline S5 (R2) Section 4.1.3. Study for effects on embryo-fetal development [ICH S5 (R2), 2005]. In short, three groups of pregnant animals are treated with the test item during organogenesis to evaluate the effect of in utero exposure on embryo-fetal development. A fourth group serves as vehicle control. The ICH guideline describes "*For all but the rarest events (such as malformations, abortions, total litter loss), evaluation of between 16 to 20 litters for rodents and rabbits tends to provide a degree of consistency between studies. Below 16 litters per evaluation, between study results become inconsistent, above 20 -24 litters per group consistency and precision is not greatly enhanced".* It is common practice to use a group size of 22 mated rats to assure 16-20 litters per group. For rats, this would result in an animal usage of 88 females for the F_0-generation and, according to European animal welfare rules, approximately 1056 fetuses for the F_1-generation (assuming a litter size of 12 fetuses/litter) per EFD study. It should be realized that these numbers are without the addition of satellite animals for toxicokinetic (TK) assessment.

The ICH guideline S5 (R2) also provides some details on why TK information is needed in developmental and reproductive toxicity (DART) testing [ICH S5 (R2), 2005]. In the development of a new drug, it is important to know if the No Observed Adverse Effect Level (NOAEL) in an EFD study provides an adequate safety margin over the anticipated human exposure. As a rule, blood samples are collected for TK evaluation on the first and last day of treatment with up to six time points per 24 hours, and in EFD rat studies this has always been done by using satellite animals [Wise, 2009]. For each EFD study in the rat, this results in an additional 21 F_0-females and, according to European animal welfare rules, approximately 252 F_1-fetuses (assuming a litter size of 12 fetuses/litter). These numbers are based on 6 F_0-females per treatment group (as each female would be bled at 3 time points) and 3 F_0-females for the control group (as these will only be bled around the anticipated T_{max} for the test item groups; 2 time points).

Historically, satellite animals for TK have been used to circumvent possible adverse effects on the study outcome as it was thought that multiple occasions of blood sampling (usually under anesthesia) of relatively large volumes (~300 µl per time point) would impact maternal health and/or embryo-fetal development.

Microsampling is a relatively new technique for obtaining small amounts of blood (~30 µl) using an EDTA coated capillary. It was developed in 2012 at AstraZeneca [Jonsson, 2012a]. Since then, it has been validated by several organizations, pharmaceutical companies and contract research organizations [Chapman 2014, Nillson 2013]. It has already been applied to obtain repeated blood samples from adult rats, mice and dogs in general toxicity studies [Jonsson 2012a, Dillen 2014, Jonsson 2012b, Korfmacher 2015, Luo 2015, Powles-Glover 2014, Sadler 2013, Verhaeghe 2017], in juvenile rat studies [Powles-Glover 201414, Niu 2016] and even on mouse fetuses and pregnant mice [Mitchard 2017]. These data have paved the way for adopting microsampling in study animals, thereby reducing or even preventing the use of satellite animals for TK blood collection, and allowing a direct link of exposure data to treatment related findings for individual animals [Chapman 2014].

However, currently not all laboratories have implemented this method and regularly the bioanalytical method is already developed with a limit of quantification that is insufficiently sensitive. Therefore in those cases, a quick win to omit satellite animals could be established by using jugular vein blood sampling. This method does not require anesthesia and causes less discomfort to the animals in comparison to other methods (i.e. discomfort score changes from moderate to mild). In addition, multiple blood samples can be obtained from one animal making it a suitable method for sampling for TK assessment as it provides one TK profile per rat. However, a high degree of competence is required for this technique.

The present study was undertaken to investigate, if blood collection from the jugular vein would not affect health and/or embryo-fetal development, and therefore satellite animals would no longer be needed in EFD rat studies.

3.2. MATERIAL AND METHODS

3.2.1. Guidelines

The study plan was reviewed and agreed by the Laboratory Animal Welfare Officer and the Ethical Committee as required by the Dutch Act on Animal Experimentation (February 1997). The study procedures were in compliance with the ICH S5 (R2): International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline on Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility, parent guideline dated 24 June 1993 (addendum dated 9 November 2000 incorporated in November 2005), Section 4.1.3. “Study for effects on embryo-fetal development”.

3.2.2. Test animals

Sixty-six time-mated female rats (Crl:WI(Han, outbred, SPF-Quality) from Charles River Laboratories Germany were assigned to the study. Untreated primiparous females were mated at the Supplier, and arrived at Day 0 or 1 post-coitum at the Test Facility (Day 0 post-coitum was the day of successful mating; confirmed by vaginal plug). The females were 10-14 weeks old on arrival and acclimatized for at least 5 days prior to start of treatment. A health inspection was performed upon receipt of the animals. The animals were randomized one day after receipt, by computer-generated random algorithm according to body weight, with all animals within ± 25% of the mean body weight per mating day. The animals were uniquely identified by indelible ink on the tail.

3.2.3. Animal husbandry

Environmental controls for the animal room were set to maintain 18 to 24°C, a relative humidity of 40 to 70%, at least 10 air changes/hour, and a 12-hour light/12-hour dark cycle. Females were individually housed in Macrolon plastic cages (MIII type, height 18 cm). Sterilized sawdust as bedding material (Lignocel S 8-15, JRS - J.Rettenmaier & Söhne GmbH + CO. KG, Rosenberg, Germany) and paper as cage-enrichment/nesting material (Enviro-dri, Wm. Lillico & Son (Wonham Mill Ltd), Surrey, United Kingdom) were supplied. They had free access to pelleted rodent diet (SM R/M-Z from SSNIFF® Spezialdiäten GmbH, Soest, Germany) and tap-water. Evaluation for contaminants was performed for diet, water, bedding and cage-enrichment/nesting material. There were no contaminants identified that could interfere with the study.

3.2.4. Experimental groups

There were 3 groups, each consisting of 22 time-mated female rats. The animals were dosed by oral gavage from Day 6 to 17 post-coitum, inclusive, with Elix water at 5 mL/kg body weight. Group 1 animals (nos. 1-22) were not used for TK blood sampling, Group 2 animals (nos. 23-44) were used for 3 time points per occasion, and Group 3 animals (nos. 45-66) were used for 6 time points per occasion. The number of animals per group was based on the standard group size as defined in ICH S5 (R2) [ICH S5 (R2), 2005].

3.2.5. In-life observations

At least twice daily, the animals were checked for mortality/viability. Clinical signs were observed at least once daily from Day 2 post-coitum onwards up to the day prior to necropsy. The time of onset, grade and duration of any observed signs were recorded. Body weights and food consumption were determined on Days 2, 6, 9, 12, 15, 18 and 21 post-coitum.

3.2.6. Toxicokinetic blood sampling

On Days 6 and 17 post-coitum, blood was collected from all animals of Groups 2 and 3 from the jugular vein by highly trained personnel. Animals were not fasted prior to sampling and warming of the animals was not required. The neck region of the animal was shaved and kept in hyperextended position. The jugular vein appears blue in color and is found 2 to 4 mm lateral to the sternoclavicular junction. A 23G needle was inserted in the caudocephalic direction (back to front) and

blood was withdrawn slowly using a 1 mL syringe. Slight pressure with a finger was applied to stop bleeding. Approximately 0.3 mL blood samples were taken and collected into tubes containing Li-heparin as anticoagulant. The animals of Group 2 were sampled at 0.5, 2 and 8 hours post-dose, and the animals of Group 3 were sampled at 0.5, 1, 2, 4, 8 and 24 hours post-dose. Left and right sides were alternated. The samples were visually checked for quality and quantity, where after the samples were discarded.

3.2.7. Postmortem examination of the dams

On Day 21 post-coitum, all animals were sacrificed using an oxygen/carbon dioxide procedure and subsequently subjected to an external examination, followed by a thoracic and abdominal examination, with special attention being paid to the reproductive organs, and for Group 2 and 3 the sites of jugular vein blood sampling. All macroscopic abnormalities were recorded, collected and fixed in 10% buffered formalin (neutral phosphate buffered 4% formaldehyde solution, Klinipath, Duiven, The Netherlands). In addition, the jugular vein blood sampling sites (skin and underlying derma and muscle) were collected from all Group 2 and 3 animals for further examination. Each ovary and uterine horn of all animals were dissected and examined as quickly as possible to determine the number of corpora lutea, the weight of the (gravid) uterus, the number and distribution of live and dead fetuses, the number and distribution of embryo-fetal deaths (early and late resorptions), the weight and sex of each fetus, and externally visible macroscopic fetal abnormalities. In case implantations were not macroscopically visible, the uterus was stained using the Salewski technique [Salewski 1964] in order to determine any former implantation sites (Salewski staining prepared using ammonium sulphide-solution 20% (Merck, Darmstadt, Germany) and Milli-Ro water (Millipore Corporation, Bedford, USA)).

All jugular vein blood sampling sites (skin and underlying derma and muscle) of Groups 2 and 3, were processed, embedded and cut at a thickness of 2-4 micrometers. These slides were stained with haematoxylin and eosin (Klinipath, Duiven, The Netherlands), and examined by a pathologist. A peer review on the histopathology data was performed by a second pathologist.

3.2.8. Fetal pathology

External, visceral, and skeletal findings were recorded as developmental variations (alterations in anatomic structure that are considered to have no significant biological effect on animal health or body conformity and/or represent slight deviations from normal) or malformations (those structural anomalies that alter general body conformity, disrupt or interfere with normal body function, or may be incompatible with life).

<u>External:</u>

Each viable fetus was examined in detail, weighed and sexed. All live fetuses were euthanized by administration of approximately 0.05 mL (=10mg) of sodium pentobarbital (Euthasol® 20%; AST Farma B.V., Oudewater) into the oral cavity. For late resorptions a gross external examination was performed.

Visceral (Internal):

Approximately 50% of the fetuses in each litter was examined for visceral anomalies by dissection in the fresh (non-fixed) state. The thoracic and abdominal cavities were opened and dissected using a technique described by Stuckhardt and Poppe [Stuckhardt 1984]. This examination included the heart and major vessels. Fetal kidneys were examined and graded for renal papillae development as described by Woo and Hoar [Woo 1972]. The sex of all fetuses was confirmed by internal examination. The heads were removed from approximately 50% of the fetuses in each litter and placed in Bouin's solution (Klinipath, Duiven, The Netherlands) for soft-tissue examination of all groups using the Wilson sectioning technique [Wilson 1965].

Skeletal:

From the other 50% of the fetuses in each litter, the sex was confirmed by internal examination. All fetuses were eviscerated, fixed in 96% aqueous ethanol, macerated in potassium hydroxide (Merck, Darmstadt, Germany) and stained with Alizarin Red S (Klinipath, Duiven, The Netherlands) by a method similar to that described by Dawson [Dawson 1926]. Skeletal examination was done for 50% of the fetuses (i.e. the fetuses with heads).

3.2.9. Data analysis

Calculations

The fetal developmental findings were summarized by: 1) presenting the incidence of a given finding both as the number of fetuses and the number of litters available for examination in the group; and 2) considering the litter as the basic unit for comparison, calculating the number of affected fetuses as a mean litter proportion on a total group basis, where:

$$\text{Viable fetuses affected/litter (\%)} = \frac{\text{number of viable fetuses affected/litter}}{\text{number of viable fetuses/litter}} \times 100$$

Statistical Analyses

All tests were two-sided and in all cases $p < 0.05$ was accepted as the lowest level of significance. If the variables could be assumed to follow a normal distribution, the Dunnett-test [Dunnett 1955] (many-to-one t-test) based on a pooled variance estimate was applied for the comparison of the treated groups and the control group. The Steel-test [Miller 1981] (many-to-one rank test) was applied if the data could not be assumed to follow a normal distribution. The Fisher Exact-test [Fisher 1950] was applied to frequency data. The Mann Whitney test [Mann 1947] was used to compare mean litter proportions (percent of litter) of the number of viable and dead fetuses, early and late resorptions, total resorptions, pre- and post-implantation loss, and sex distribution. Mean litter proportions (percent per litter) of total fetal malformations and developmental variations (external, visceral and skeletal), and each particular external, visceral and skeletal malformation or variation were subjected to the Kruskal-Wallis nonparametric ANOVA test [Kruskal 1952] to determine intergroup differences. If the ANOVA revealed statistically significant ($p<0.05$) intergroup variance, Dunn's test [Dunn 1964] was used to compare the compound-treated groups to the control group.

Definitions
Fetal (late) resorptions were defined as a dead fetus with external degenerative changes and presence of distinguishable features such as head or limbs. Embryonic (early) resorptions were defined as evidence of implantation without presence of distinguishable features such as head or limbs. Dead fetus was defined as a non-viable fetus without external degenerative changes and presence of distinguishable features such as head or limbs. Post-implantation loss included embryonic (early) resorptions, fetal (late) resorptions and dead fetuses.

3.3. RESULTS

3.3.1. Maternal findings

Mortality
No mortality occurred during the study period.

Clinical Signs
No treatment-related clinical signs were noted during the observation period.
Alopecia was noted for one animal (no. 60) in Group 3 from Day 16 post-coitum onwards. This finding occurred within the range of background findings to be expected for rats of this age and strain which are housed and treated under the conditions in this study, and therefore was considered not treatment-related.

Body Weights and Food Consumption
No treatment-related changes in body weights, body weight gain or body weight gain corrected for gravid uterus were noted. In addition, food consumption before or after correction for body weight was not affected by blood sampling (see Figure 3.1).

Macroscopic Examination
Necropsy did not reveal any treatment-related alterations.
The finding of alopecia for animal no. 60 was confirmed at macroscopic examination. No further findings were noted.

Microscopic Examination
Procedure-related microscopic findings were noted at the jugular vein blood sampling sites (skin and underlying derma and muscle) of the females of Groups 2 and 3 and are summarized in Table 3.1.

In general, recorded findings at the jugular vein blood collection sites were mild in nature in both groups and showed a slight increase in incidence and/or severity in Group 3. Only the severity of hemorrhages increased up to moderate degrees after doubling the sampling frequency, but this was still considered acceptable.

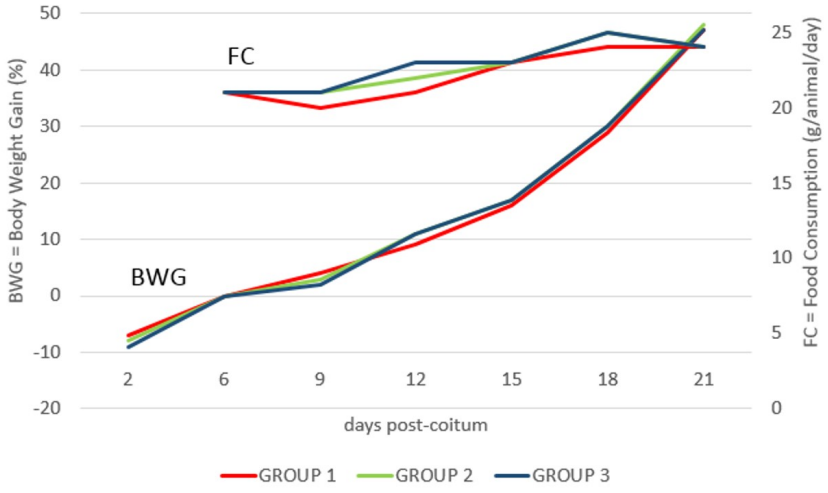


Figure 3.1. Body weight gain (%) and food consumption (g/animal/day) of F_0-females during post coitum. Group 1 (control), Group 2 (3 time points of blood collection) and Group 3 (6 time points of blood collection). None were significantly different from the control group.

One Group 3 female showed slightly more severe lesions at the dermis/subcutis of the sampling site, consisting of moderate hemorrhage and fibroplasia, and slight inflammation and necrosis.

Maternal Pregnancy Data

The numbers of pregnant females, corpora lutea and implantation sites, and pre-implantation loss in the control and test groups were similar and in the range of normal biological variation. See Table 3.2.

One female in Group 2 (no. 35) was not pregnant.

Table 3.1. Microscopic findings.

	Females	
Blood sampling frequency on Day 6 and 17 post-coitum:	Three times (Group 2)	Six times (Group 3)
SKIN/SUBCUTIS [a]	22	22
DERMIS/SUBCUTIS BLOOD VESSEL		
Degeneration vascular wall		
Minimal	1	1
Fibrinoid necrosis vascular wall		
Minimal	2	2
Slight	1	3
Thrombus		
Minimal	1	-
EPIDERMIS		
Pustule(s)		
Minimal	3	3
DERMIS/SUBCUTIS INCLUDING UNDERLYING MUSCLE		
Hemorrhage dermis/subcutis		
Minimal	11	8
Slight	3	9
Moderate	-	4
Inflammation dermis/subcutis		
Minimal	10	15
Slight	1	1
Necrosis dermis/subcutis		
Slight	-	1
Fibroplasia dermis/subcutis		
Minimal	3	6
Slight	-	2
Moderate	-	1

[a] = Number of tissues examined from each group.

3.3.2. Fetal findings

For further detail on summary data of fetal findings, see Table 3.2.

Table 3.2. Summary of fetal data.

Group		Sex		Viable	Dead	Resorptions		Post Implantation	Implantation	Corpora	Pre Implantation	Fetal Weights	No. of Gravid
		M	F	Fetuses	Fetuses	Early	Late	Loss	Sites	Lutea	Loss	(gram)	Females
1	TOTAL	125	130	255	0	9	1	10	265	277	12	NA	22
	MEAN	5.7	5.9	11.6	0.0	0.4	0.0	0.5	12.0	12.6	0.5	5.3	
	S.D.	2.21	1.77	2.11	0.00	0.59	0.21	0.67	1.94	1.71	0.96	0.29	
2	TOTAL	117	106	223	0	11	0	11	234	271	37	NA	21
	MEAN	5.6	5.0	10.6	0.0	0.5	0.0	0.5	11.1	12.9	1.8	5.4	
	S.D.	1.66	2.01	2.25	0.00	0.87	0.00	0.87	1.82	1.92	1.79	0.29	
3	TOTAL	122	132	254	0	16	0	16	270	290	20	NA	22
	MEAN	5.5	6.0	11.5	0.0	0.7	0.0	0.7	12.3	13.2	0.9	5.4	
	S.D.	1.97	2.00	2.32	0.00	1.35	0.00	1.35	2.23	1.71	1.19	0.29	

Group 1 (control), Group 2 (3 time points) and Group 3 (6 time points). No statistically significant changes were observed compared to the control group.
NA = not applicable

Litter Size

There were no treatment-related effects on litter size of any group.
Mean litter sizes were 11.6, 10.6 and 11.5 viable fetuses/litter for Group 1 (control), Group 2 (three time points) and Group 3 (six time points), respectively.

Sex Ratio

The male:female ratio was unaffected by treatment.
Mean sex ratios (males:females) were 48:52, 53:47 and 48:52 for Group 1 (control), Group 2 (three time points) and Group 3 (six time points), respectively.

Fetal Body Weight

There were no effects on fetal body weights (both sexes) noted by treatment.
Mean combined (male and female) fetal body weights were 5.3, 5.4 and 5.4 gram for Group 1 (control), Group 2 (three time points) and Group 3 (six time points), respectively.

3.3.3. Fetal morphological examinations

The numbers of fetuses (litters) available for morphological examination were, 255 (22), 223 (21), and 254 (22) in Groups 1, 2, and 3, respectively. External examination was performed for all fetuses, visceral examination was performed for approximately half of the fetuses of all groups, and skeletal examination was done for the other half of fetuses.

External Malformations and Variations

There were no treatment-related effects on external morphology. External malformations occurred in Group 1 in one fetus and in Group 2 in two fetuses. One Group 2 fetus (of dam 31) had an exophthalmos combined with an absent lower jaw. A small lower jaw was seen in one control fetus (of dam 16). One Group 2 fetus (of dam 25) had inwards rotated hind limbs. These findings were confirmed by skeletal examination. The single occurrence and group distribution of the above malformations do not indicate a treatment relationship and therefore all were considered to be coincidental findings. There were no external variations.

Visceral Malformations and Variations

There were no treatment-related effects on visceral morphology. One Group 2 fetus (of dam 31) appeared to have two visceral malformations, situs inversus and a cyst attached to the pancreas. As this was a single occurrence, it was considered a coincidental finding. No other visceral malformations were noted. Visceral variations were observed in 14.0%, 5.5% and 6.7% of fetuses per litter in Groups 1, 2, and 3, respectively. These all occurred in the absence of a treatment-related incidence, infrequently and/or at frequencies that were within the range of available historical control data.

Skeletal Malformations and Variations

There were no treatment-related effects on skeletal morphology. Skeletal malformations were observed in three fetuses of Group 1, one fetus of Group 2 and three fetuses of Group 3. In the control group, one fetus (of dam 2) had a rib anomaly, one fetus (of dam 15) was assigned with a costal cartilage anomaly, and one fetus (of dam 21) had a sternoschisis of sternebra 6. One Group 2 fetus (of dam 31) had besides skeletally confirmed absent mandibles, also a vertebral anomaly. In Group 3, one fetus (of dam 46) had fused skull bones combined with a vertebral anomaly, sternum anomaly and bent scapulae, one fetus (of dam 50) also showed a costal cartilage anomaly, and one fetus (of dam 52) had a cervical vertebral anomaly. The group distribution and single occurrence of these malformations do not suggest any treatment relationship and therefore all were considered to be coincidental findings. Skeletal variations occurred at an incidence of 74.9 %, 74.5% and 74.4% per litter in Groups 1, 2 and 3, respectively. These were not considered treatment-related as they occurred in the control group only, infrequently and/or at frequencies that were within the range of available historical control data.

For further detail on summary data of fetal morphological findings, see Table 3.3 for malformations and Table 3.4 for variations.

Table 3.3. Summary of fetuses and litters with malformations (absolute numbers).

	No. of fetuses per group			**No. of litters per group**		
	1	2	3	1	2	3
NUMBER EXAMINED EXTERNALLY	255	223	254	22	21	22
Limb(s), malrotated	0	1	0	0	1	0
Exophthalmos	0	1	0	0	1	0
Lower jaw, absent or small	1	1	0	1	1	0
NUMBER EXAMINED VISCERALLY	127	110	127	22	21	22
Situs inversus	0	1	0	0	1	0
Viscera, cyst(s)	0	1	0	0	1	0
NUMBER EXAMINED SKELETALLY	129	115	127	22	21	22
Rib anomaly	1	0	0	1	0	0
Sternoschisis	1	0	0	1	0	0
Costal cartilate anomaly	1	0	1	1	0	1
Skull anomaly	0	0	1	0	0	1
Vertebral anomaly	0	1	2	0	1	2
Sternum anomaly	0	0	1	0	0	1
Bent limb bone(s)	0	0	1	0	0	1

Group 1 (control), Group 2 (3 time points) and Group 3 (6 time points). No statistically significant changes were observed compared to the control group.

Table 3.4. Summary of fetuses and litters with variations (absolute numbers).

	No. of fetuses per group			No. of litters per group		
	1	2	3	1	2	3
NUMBER EXAMINED EXTERNALLY	255	223	254	22	21	22
Number of litters with findings	0	0	0	0	0	0
NUMBER EXAMINED VISCERALLY	127	110	127	22	21	22
Liver, small supernumerary lobe(s)	10	3	7	7	3	5
Liver, appendix	6	3	1	6	3	1
Ureter(s), dilated	1	0	0	1	0	0
Aortic arch, supernumerary artery	0	0	1	0	0	1
NUMBER EXAMINED SKELETALLY	129	115	127	22	21	22
Sternebra(e) malaligned (slight or moderate)	27	28	28	18	16	18
Bent rib(s)	10	15	4	7	10	4
14th Rudimentary rib(s)	62	68	71	21	21	20
Vertebral centra, reduced ossification	4	0	1	4	0	1
7th Cervical ossification site(s)	4	7	7	4	5	6
Reduced ossification of the skull	7	5	1	4	4	1
14th Full rib(s)	14	2	7	8	2	6
Pelvic girdle, caudal shift	7	6	6	4	6	4
Metacarpal(s) and/or metatarsal(s) unossified	4	3	3	2	3	3
Supernumerary sternebra	2	0	0	1	0	0
Rib(s), reduced ossification	1	0	0	1	0	0
Vertebral arches, reduced ossification	1	0	0	1	0	0
Sternebra(e), branched	0	0	2	0	0	2
7th Cervical full rib(s)	0	0	2	0	0	2
Sternebra(e) #5 and/or #6 unossified	0	0	1	0	0	1

Group 1 (control), Group 2 (3 time points) and Group 3 (6 time points). No statistically significant changes were observed compared to the control group.

3.4. DISCUSSION AND CONCLUSIONS

Mated female Wistar Han rats were assigned to three dose groups, each containing twenty-two animals. Water was administered once daily by oral gavage from Day 6 to 17 post-coitum. On the first and last day of dosing (i.e. Day 6 and 17 post-coitum) blood was sampled from the jugular vein of the animals of Groups 2 and 3. On both occasions, blood was sampled from Group 2 animals three times after dosing and from Group 3 animals six times after dosing.

When implementing jugular vein blood sampling, a high degree of competence is required for this technique and sufficient time should be allowed to fully train the technicians. Correct restraint of the animal and precise insertion of the needle is crucial for successful blood sampling and minimal animal discomfort.

The current study did not show any adverse maternal or developmental effects for both groups (i.e. 3 and 6 blood sampling time points). Only very mild, and hence acceptable histologic changes at the jugular vein blood sampling sites were noted, whereby increased frequency of sampling resulted in a subtle increase in incidence and/or severity of most of these findings. There was one exception in a Group 3 female (6 times sampled) with slightly more severe lesions. However, at this single occurrence it was considered not to impact the overall evaluation.

In conclusion, these results strongly suggest that satellite animals will no longer be needed in EFD studies in the rat when TK blood samples are collected from the jugular vein, and will hereby reduce the number of animals in a standard EFD studies for pharmaceuticals by 20% (21 F_0-females and, according to European animal welfare rules, approximately 252 F_1-fetuses). This method of blood sampling would therefore be recommend in case microsampling cannot be used. Moreover, similar to microsampling, it may decrease variability due to calculation of TK parameters from composite samples, and would allow direct comparison of individual effects to individual exposures.

One aspect that should be emphasized, is the impact of excluding the use of satellite animals on the assessment of fetal exposure. In case fetal exposure data is requested, maternal animals should be dosed beyond the standard dosing period (until Day 20 post-coitum instead of Day 17 post-coitum). Extension of the dosing period would omit the standard post-dosing recovery period of 3 days which could result in higher toxicity and a lower NOAEL. In rare cases it might underestimate the incidence of fetal malformations as more malformed fetuses may die with prolonged maternal dosing, and resorption rates could be increased. Fetal malformation rates will not increase because these are induced during the period of major organogenesis (i.e. from Days 6-17 post-coitum), included in both dosing regimens, except for late-induced malformations (e.g. premature closure of the ductus arteriosus). On the other hand, by omitting the treatment-free recovery period it would be a better model for human exposure which is of course the ultimate reason for performing the EFD study. Since numerous organ systems continue to develop in the late phase of the fetal period, continued exposure during this stage would enable a complete assessment of developmental toxicity. Moreover, the true added value of fetal exposure assessment should be critically appraised. Neither the ICH S5(R2) nor ICH S3A guidance documents [ICH S5(R2) 2005, ICH S3A 1994] specifically recommend assessment of embryonic/fetal exposure, but some pharmaceutical companies would request this by default. Maternal exposures provide information relevant to human exposure, however relevant human fetal data are hardly available. As fetal exposures are determined late in gestation (when sufficient fetal blood can be obtained), the rel-

evance of exposure at this time to earlier developmental stages is ambiguous. Therefore the goal to collect these data should be carefully considered with regard to estimating human risk [Wise 2009, Wier 2006, Brannen 2011].

When hematology and/or clinical biochemistry parameters are required in the EFD rat study (based on target toxicity), the addition of satellite animals would be recommended in case microsampling cannot be used because the total blood sample volume collected for TK would impact these parameters. Moreover, frequent blood sampling could make rats more vulnerable to the effects of drugs with hemodynamic actions on the circulatory system or on hematopoiesis. On the NC3Rs (National Centre for the Replacement, Refinement & Reduction of Animals in Research) website a decision tree is published which can help to determine if your study is amenable to microsampling, and which is the most relevant approach [NC3Rs 2017].

A case for microsampling:
It is important to know that the use of microsampling would not only be advisable in above example, but in many more. For animal welfare reasons, implementation of microsampling for all type of studies and all animal species (i.e. small and large) will be a relatively simple but crucial step in respect of the 3Rs. As microsampling can only be used in case the bioanalytical method is validated for these small quantities, it is essential that at an early stage in the development of the bioanalytical method it is investigated if it is sensitive enough to be used in combination with microsampling. It would therefore be important to get more awareness for this in a broader field of animal testing, and to encourage all bioanalytical laboratories to develop methods using microsamples.

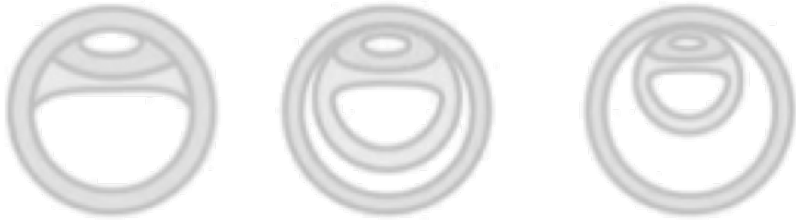
With the ongoing improvements of the current methods for DART testing, all with the aim to optimally implement the 3R principles, it will be important for the involved scientists in all sectors (i.e. government, public interest, industry, university, and consulting laboratories) to engage and keep each other up-to-date on new developments. This paper focuses on TK assessment of pharmaceuticals, however with the global endeavor to optimize testing strategies and implement the 3Rs, these results can also be used for DART testing of other type of compounds like agrochemicals and industrial chemicals. In case of absent toxicity at the limit dose of 1000 mg/kg body weight/day, TK data will provide valuable information if the animal model was adequately exposed. Moreover, use of TK can aid to ensure studies are designed to be relevant for risk assessment in humans, and circumvent using disproportionately high doses that could result in preventable suffering in experimental animals, which is of great importance.

Conflict of interest
The authors declare that there are no conflicts of interest.

Acknowledgements
The authors would like to thank the involved personnel of Charles River ‘s-Hertogenbosch for their highly qualified and dedicated work.

Chapter 4

The underestimated value of OECD 421 and 422 repro screening studies: putting it in the right perspective



Manon Beekhuijzen[a], Michiel Alexander de Raaf[b], Amy Zmarowski[a], Francois van Otterdijk[a], Birgit Peter[a], Harry Emmen[a]

[a] WIL Research Europe B.V., Hambakenwetering 7, 5231 DD, 's-Hertogenbosch, The Netherlands

[b] Department of Pulmonology, VU University Medical Center / Institute for Cardiovascular Research, De Boelelaan 1117, 1081 HZ, Amsterdam, The Netherlands

Published in: Reproductive Toxicology. 2014 Sep;48:81-7. doi: 10.1016/j.reprotox.2014.04.003. Epub 2014 Apr 21.

ABSTRACT

To assess the efficacy of reproduction/developmental screening studies (OECD 421 and 422), a retrospective evaluation of 134 studies was performed. The major findings were: (1) for up to half of the studies with developmental and reproductive toxicity, these effects would have been missed in other types of studies, which underscores that reproduction/developmental screening studies should not be waived by default based on negative 28-day and/or prenatal developmental data, (2) the required number of animals as stated in the guidelines, is appropriate for detecting developmental and reproductive toxicity, and (3) adding measurements like anogenital distance, internal sex determination and nipple retention, plus extending the postnatal period would add predictive value. Overall, the current reproduction/developmental screening studies are effective in providing unique data, especially considering the limited number of animals used. Some simple additions would enrich its value in risk assessment even further.

Keywords
Reproduction/developmental screening test, OECD 421, OECD 422

4.1. INTRODUCTION

According to the European REACH legislation (REACH), conduct of a reproduction/developmental screening test (OECD 421 [OECD 421] or 422 [OECD 422]; further referred to as 'repro screening study') is required for industrial chemicals with an annual production or import level of 10 tonnes or more. The OECD test guideline 421 describes the 'Reproduction/Developmental Toxicity Screening Test' and the OECD test guideline 422 is the 'Combined Repeated Dose Toxicity Study with Reproduction/Developmental Toxicity Screening Test' (which is the OECD 421 combined with a 28-day toxicity study; OECD 407 [OECD 407]). The OECD 421 and 422 correspond to the US-EPA Health Effects Test Guidelines OPPTS 870.3550 [EPA 870.3550] and 870.3650 [EPA 870.3650], respectively. These guideline studies are generally considered as 'merely screening studies', and are described in the OECD 421 and 422 guidelines as 'neither an alternative nor is it meant to replace the existing guidelines as the prenatal developmental toxicity study (OECD 414 [OECD 414] or OPPTS 870.3700 [EPA 870.3700]) or the two-generation study (OECD 416 [OECD 416] or OPPTS 870.3800 [EPA 870.3800])'. Also, the value of repro screening studies is debated in literature. For example, Aulmann [Aulmann 2012] suggests that information provided from a 28-day and prenatal developmental study is sufficient to waive the repro screening study. However, the practical value of repro screening studies has not been empirically evaluated by the executors of these guidelines, primarily Contract Research Organizations (CROs). The developmental and reproduction toxicity (DART) endpoints assessed in repro screening studies are essentially similar to two-generation studies [EPA 870.3800, Aulmann 2012], except for 4 differences; (1) a shorter premating period, (2) no second generation, (3) a relatively low number of animals used and (4) a limited postnatal period for pup evaluation. These differences formed the debate on the added value of a one or two-generation study (further referred to as generation study) compared to a screening study and the possible shortcomings of repro screening studies. The first two differences were already actively discussed in the literature. A two-week pre-mating period is now considered as sufficient when including detailed histopathological examination of the testes. Additionally, the necessity of a second generation was thoroughly evaluated in discussions on the extended one-generation reproductive toxicity study (EOGRTS; OECD 443 [OECD 443]) and it was concluded that classification and labelling was not impacted by eliminating the second generation [Rorije 2011].

Less information is available on the last two differences (i.e. the relatively low number of animals used per group and the limited postnatal period for pup evaluation), because for answering that question conclusively, one should compare generation studies to repro screening studies with the same substance. To perform such retrospective analyses, the current available data is too limited, although these comparisons can be conducted within the next decade as REACH will supply more data from both study types. However, it is essentially important to start shaping the right perspective in which these comparisons should be made by drawing the merits and pitfalls of the repro screening study by the executors of the guideline, as for example CROs. WIL Research Europe, 's-Hertogenbosch, The Netherlands has performed 134 repro screening studies following similar standard operating procedures between 2007 and 2013. To evaluate the efficacy of re-

pro screening studies, a retrospective analysis of DART endpoints was performed. For all repro screening studies, the incidence of developmental only, reproductive only and combined DART findings were inventoried and the following questions were addressed:

1) How distinctive are the DART findings in repro screening studies, and are these also detectable in other study types? Under which circumstances could waiving of a repro screening study be justified?
2) Is the number of animals (10/sex/group) sufficient for detecting and thoroughly evaluating any DART findings (except teratogenicity)?
3) The following limitations were assessed: could ambiguous DART findings be resolved by adjusting the study design? Was the duration of the postnatal period sufficient for adequate evaluation of potential adverse effects on pups? How should a recovery phase best be included?

4.2. MATERIALS AND METHODS

A robust evaluation was performed on 134 repro screening studies (43 OECD 421 [OECD 421] and 91 OECD 422 [OECD 422] studies) with Wistar Han rats (Crl:WI(Han), outbred, SPF-Quality from Charles River Deutschland, Sulzfeld, Germany) performed at WIL Research Europe B.V., The Netherlands, from 2007 to 2013. All studies were conducted with industrial chemicals for REACH registration purposes, followed Good Laboratory Practice (GLP) criteria, and were approved by the animal welfare committee. All sponsor- and compound-related details remained confidential. Prior to inclusion, all studies were re-evaluated to assure the following assessments were harmonized: the validity according to dose levels, determination of reproductive and developmental findings and the threshold to whether a finding was considered related or unrelated to the given treatment. As an insufficient dose level would result in no findings, this would bias this evaluation in respect to its efficacy. All studies were verified to have a NOAEL below the highest dose level tested, or a highest dose level of 1000 mg/kg. For a few studies, these two criteria were not met; dose levels were however still considered valid because slightly higher dose levels tested in the dose range finding study produced severe toxicity, which indicated at steep dose response curve for these compounds. Reproduction findings were defined as effects in the phase up to and including implantation. From implantation onwards, changes were defined as developmental effects. Reproductive parameters investigated in these studies were mating, fertility and conception indices, precoital time, and numbers of corpora lutea and implantation sites. Developmental parameters included gestation index and duration, parturition, maternal care and early postnatal pup development (consisting of mortality, clinical signs, body weight and macroscopy). When performing a repro screening study in a CRO setting, interpretation whether a finding is treatment related is evaluated in respect to the nature and severity of the findings, the number of affected animals, the incidence in historical control data, any corroborative findings and a possible dose response relationship. For example, the incidence of non-pregnant females was ~10% in the control groups of these 134 studies, which underlines the sensitivity of the study design versus the

biological variation. For all projects, the following details were included in an in-house developed Microsoft Access database:

- For each assessed parameter, the number of affected animals or litters per group.
- The number of pregnant females and litters obtained per group.
- Effects on reproduction and/or development: clear reproductive toxicity, clear developmental toxicity, no effects on reproduction or development, or ambiguous effects.
- If and how a recovery phase was added.

An inventory on the incidence of developmental, reproductive and combined DART findings was made and queries were then performed for answering the following objectives:

4.2.1. Type of DART findings

Presently, a general approach is to waive a repro screening study based on information obtained in a 28-day toxicity study (OECD 407 [OECD 407]) and/or prenatal developmental study (OECD 414 [OECD 414]). Aulmann [Aulmann 2012] stated that the information obtained from a combination of these two studies is considered more robust and relevant for the assessment of reproductive toxicity than the information obtained from the OECD 421/422 screening studies only. To investigate this, all studies with developmental and/or reproductive toxicity were studied in detail and analyzed regarding the number of studies in which the DART findings were confined to parameters that are also investigated in prenatal developmental studies and/or 28-day repeated dose toxicity studies.

4.2.2. Sufficient number of animals per group

The repro screening guidelines mention 'it is recommended to start with at least 10 animals per sex and would result in at least 8 pregnant females per group'. Failure to achieve the desired number of pregnant animals does not necessarily invalidate the study and should be evaluated on a case-by-case basis, considering a possible causal relationship to the test substance. The guidelines however do not mention the number of litters required for a thorough evaluation of developmental effects. In the present evaluation, the number of groups was counted in which less than 8 pregnant females or litters were obtained. Additionally, for those studies having insufficient pregnancies and/or litters, the cause of this was evaluated as: a treatment related effect on reproductions and/or development, (severe) maternal toxicity, a chance finding for control, low dose or mid dose groups with an unaffected high dose group, or a chance finding for high dose group. For example, a finding was considered to have occurred by chance as it was caused by single unrelated occurrences (e.g. one female did not mate, one female was not pregnant, and one female died before mating) with the remaining litters being unaffected. To evaluate if the study design with this number of animals per group is sufficient for detecting DART findings in these screening studies, the percentage of affected animals and litters was calculated for studies that showed treatment related reproductive and/or developmental toxicity.

4.2.3. Limitations of repro screening studies

The following three items were assessed:

Ambiguous study results

The frequency of ambiguous DART findings was determined to assess the robustness of the study design. An effect was considered ambiguous when a single parameter was affected (no corroborative findings), it was a borderline finding (e.g. just outside normal limits) and only the high dose group was affected. In case of ambiguous findings, it was examined if these could be clarified by having additional animals and/or parameters included in the study design.

Prolongation of postnatal period

Based on the obtained data, it was evaluated if the relatively short postnatal period (4 days) and few pup evaluations were too limited for an adequate evaluation of pup toxicity. The database was inspected for any studies in which a longer postnatal period or additional pup parameters would have been beneficial for interpretation.

Addition of recovery phase

The OECD 422 guideline mentions that when examining the potential for reversibility of effects (as can be part of the OECD 407 study), recovery females must be non-mated. A major disadvantage of this approach is that data from non-mated recovery females are compared with results from lactating main females; absence of toxicity could mean a possible recovery but could also be due to the difference in physiological state between the never-pregnant and lactating females. To circumvent this, two additional study designs can be used: (1) assess recovery in males only (this would add 10 additional animals; 5 males each in control and high dose groups) or (2) add recovery males and non-mated recovery females (5/sex in control and high dose groups) together with non-mated main females (5 animals/group); this would result in forty additional animals, thus using 50% more animals. Studies with a recovery phase were evaluated to determine if the three types of recovery designs could impact the evaluation. The pros and cons were evaluated.

4.3. RESULTS

Out of the 134 repro screening studies, 103 studies (77%) did not show any effect on reproduction or development, 27 studies (20%) showed clear reproductive and/or developmental toxicity and 4 studies (3%) had ambiguous DART findings (Figure 4.1). The similarity in endpoints between the OECD 421 and OECD 422 studies is confirmed by the little differences found in the incidence of DART findings.

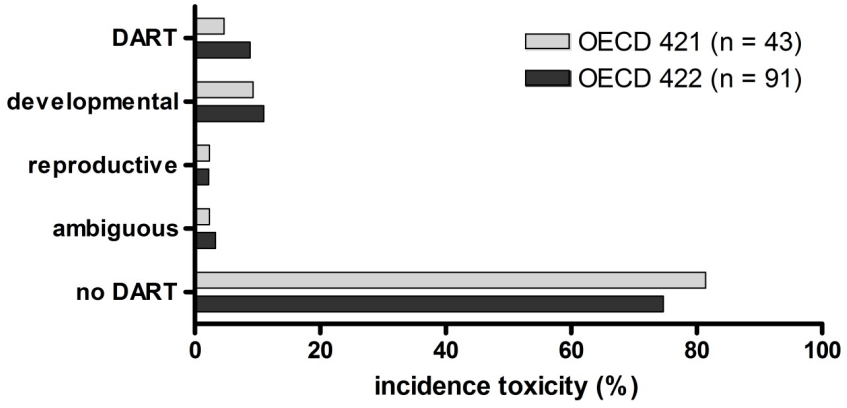


Figure 4.1. Percentage of repro screening studies (OECD 421 and 422) with developmental and/or reproductive toxicity (x-axis). The studies were classed in having DART findings, developmental findings, reproductive findings, ambiguous findings or no DART findings (y-axis).

4.3.1. Type of DART findings

In our dataset, the type of DART findings were assessed and evaluated whether these affected parameters could be detected in the 28-day repeated dose toxicity study (OECD 407), prenatal developmental study (OECD 414) or solely in a reproduction study (Figure 4.2).

Out of the 27 studies that showed clear treatment related reproductive and/or developmental effects, there were 7 studies in which the findings would also be detected in a standard 28-day repeated dose toxicity study (OECD 407) in which assessment of sex organs is included by default. Of these, six studies had findings on the male sex organs and one study showed effects on the female mammary gland which explained the increased occurrence of postnatal pup mortality (the pups had a concurrent absence of milk in the stomach).

In 10 cases of the 27, the type of developmental effects may have been seen when performing a prenatal developmental study (OECD 414). Findings included decreased pup viability, decreased pup body weights at birth and/or an affected sex ratio. For 3 of these cases, only pup viability was affected. As the cause of death for these pups could not be determined (no abnormalities were noted), it could be attributable to something other than a fatal malformation (e.g. a postnatal manifestation or maternal effects), and might have been undetected in developmental toxicity studies. In that instance, only 7 cases could be seen in developmental toxicity studies.

For the remaining 10 (out of 27) studies with DART effects (or 13 studies when including the studies with undetermined postnatal death), toxicity consisted of reduced numbers of corpora lutea and implantation sites, decreased fertility, conception and gestation indices, increased precoital time, and/or compromised maternal care. These findings would only be detected in reproduction studies.

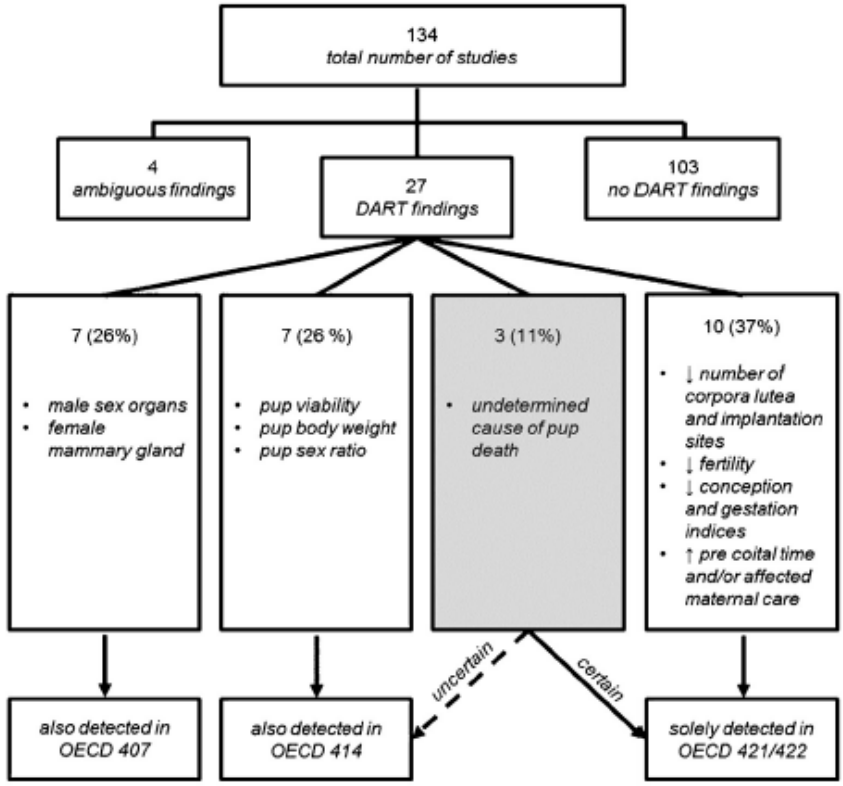


Figure 4.2. Breakdown of the number of studies (%) and specifications of DART findings: solely detected in the repro screening study (OECD 421/422), could also be found in the 28-day repeated dose toxicity study (OECD 407) or prenatal developmental study (OECD 414) when testing the same test compound.

The occurrence of false-negative findings in the 103 studies without DART effects could, in theory, not be revealed thoroughly, as the DART endpoints in the OECD 407 or 414 are limited or different. In addition, for all OECD 421 studies without DART findings (N = 36), the OECD 407 studies which were done with the same test compound, confirmed absence of findings on reproductive organs. This evaluation was summarized in Figure 4.3.

Out of the 27 studies that showed clear treatment related reproductive and/or developmental effects, there were 7 studies of which the findings would be reflected in a standard 28-day repeated dose toxicity study (OECD 407) in which assessment of sex organs is included by default. Six studies with findings on the male sex organs and one study showed effects on the female mammary gland which explained the increased occurrence of postnatal pup mortality (with concurrent absence of milk in the stomach).

In 10 cases, the type of developmental effects may have been seen when performing a prenatal developmental study (OECD 414). Findings included decreased pup viability (probably due to birth defects), decreased pup body weights and/or affected sex ratio. For 3 of these cases, only pup viability was affected. As the cause of death for these pups could not be determined (no abnormalities were noted), it could be attributable to something other than a fatal malformation (e.g. a postnatal manifestation), and might have been undetected in developmental toxicity studies. In that instance, only 7 cases could be seen in developmental toxicity studies.

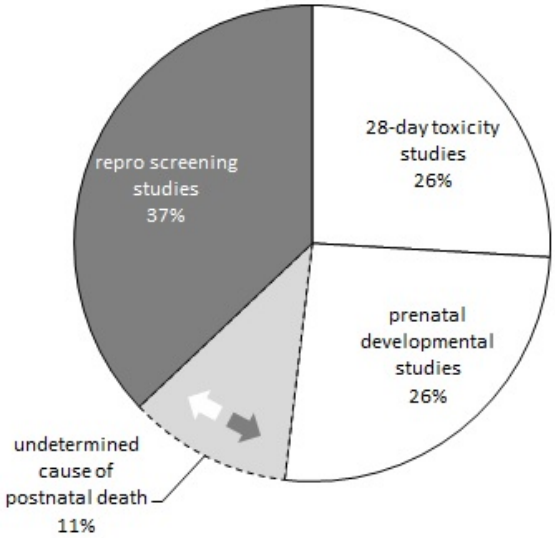


Figure 4.3. Subdivision (incidence of studies) based on the type of DART findings that would also be noted in 28-day toxicity studies (26%), prenatal developmental studies (26%), or solely in repro screening studies (37%). For the studies indicating 'undetermined cause of postnatal death' (i.e. studies in which it is unclear if postnatal deaths resulted from prenatal abnormalities), the 11% (3 studies) could either be evident in prenatal developmental studies (37%) or solely in repro screening studies (48%).

For the remaining 10 studies with DART effects (or 13 studies when including the studies with undetermined postnatal death), toxicity consisted of reduced numbers of corpora lutea and implantation sites, decreased fertility, conception and gestation indices, increased precoital time, and/or affected maternal care. These findings would only be detected in reproduction studies.
This evaluation was summarized in Figure 4.3.

4.3.2. Sufficient number of animals per group

When using 10 animals/sex/group, out of 134 screening studies (i.e. 536 groups), 29 studies had less than 8 pregnant females in at least one group (a total of 33 groups were affected). Thirteen were due to treatment related developmental and/or reproductive toxicity; this included one study with 3 non-pregnant females in the high dose group that was considered ambiguous, which

is evaluated in Section 4.3.3. Eleven studies had insufficient pregnancies, which was considered attributable to chance as the high dose group was not affected. Four had (severe) maternal toxicity that resulted in a reduced number of pregnant females. In the remaining study, the slightly lower number of pregnant females in the high dose group was considered attributable to a chance finding as of the 10 paired females, only 9 mated and 7 of those were pregnant (Figure 4.4, left panel). In summary, out of the 29 studies with insufficient pregnant females, one of these gave ambiguous results because of this. Out of the 536 groups in 134 screening studies, 50 groups in 42 studies had less than eight litters. This is 13 more than the 29 studies with 'insufficient pregnancies in at least one group'. Six additional cases were due to treatment related DART findings, 4 additional cases were caused by (severe) maternal toxicity, and the 3 remaining cases were considered chance findings (Figure 4.4, right panel).

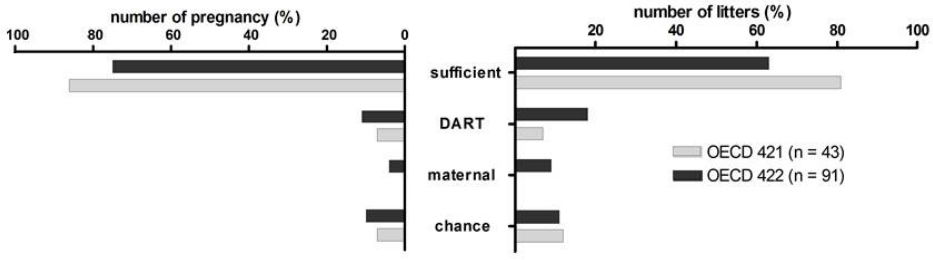


Figure 4.4. Percentage of repro screening studies in which (in)sufficient pregnant females and litters were obtained with subdivision into the reason for these lower numbers. The left panel reports the % of studies with sufficient number of pregnant females (top row), % of studies with insufficient pregnant females due to DART (second row), due to maternal toxicity (third row) or by chance (bottom row). The right panel shows the same rows but then in respect to the number of litters per group.

In 27 repro screening studies with DART findings, the average incidence of DART findings was 79% (median incidence was 94%). See Figure 4.5. At least 38% of the animals or litters per group were affected.

Apart from the 27 studies with clear DART findings, there were 4 studies with ambiguous effects. From these 4 studies, 2 cases were considered equivocal because of the relatively low percentage of animals affected (i.e. 22 and 30%). The other 2 studies had an incidence of 71% and 78%. For details see Section 4.3.3.

4.3.3. Limitations of repro screening studies

Ambiguous study results

For 4 studies (i.e. 3%; three OECD 422 and one OECD 421), ambiguous DART effects were noted. For one study, only 7 females at the high dose group were pregnant and no other effects were noted. Therefore, this slightly low fertility and conception rate may have been a chance finding but a possible effect on reproduction could not be disregarded. One study had 2 females in the high dose group

(out of 9mated females) with relatively small litter sizes (i.e. 3 and 6 pups per litter). The remaining 7 litters were of normal size. The other 2 projects were considered ambiguous as the only parameter affected was the sex ratio (with normal sized litters). In one of these studies, the high dose group showed an overall sex ratio of 38:62 (percent of males:females), and in 7 out of 9 litters more female pups were born. In the other study, the high dose group showed an overall sex ratio of 64:36 (percent of males:females), and 5 out of 7 litters consisted of more male than female pups. Therefore we assessed the sex ratio of all 134 studies, which is evaluated in the discussion.

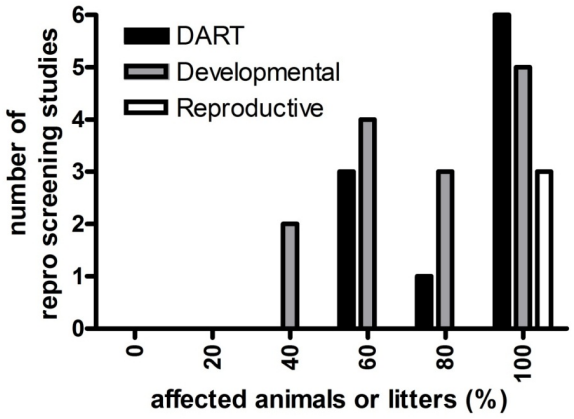


Figure 4.5. Number of repro screening studies with DART findings classified by percentage (in 20% bins) of affected animals or litters per group with developmental toxicity, reproductive toxicity or both (DART).

Prolongation of postnatal period

None of these studies had a prolonged postnatal period and therefore a retrospective assessment on the added value of this could not be evaluated. To give insight on this matter, an additional evaluation was performed using in-house obtained one- and two-generation data. See Section 4.4 for details.

Addition of recovery phase

Fifteen of the 91 OECD 422 studies included a recovery phase. Out of these, 7 studies were performed with recovery males only, 7 studies with males and non-mated recovery females, and 1 study with males, non-mated main females and non-mated recovery females. For one of these projects (performed according to the guideline with non-mated recovery females), recovery from general toxicity could not be correctly evaluated. Namely, decreased hemopoietic foci in the spleen with concurrent changes in haematology parameters were noted for mated high dose main females. These findings were however not seen in the nulliparous recovery females. Normally, pregnancy leads to an increase in hemopoietic foci in the spleen. The hemopoietic foci seen in main high dose females more closely resembled the amount present in nulliparous females, which may indicate the normal increase in hemopoietic foci seen at pregnancy is blocked by the test item. Since the recovery animals were not mated, the possible recovery from this could not be investigated.

4.4. DISCUSSION

A thorough retrospective evaluation of 134 repro screening studies (OECD 421 and 422 [OECD 421, OECD 422]) was performed to investigate the efficacy and sensitivity of these types of studies, and to place the value of repro screening studies in the right perspective. The major findings were: (1) up to half of the studies with findings are solely found in repro screening studies and not reflected in 28-day toxicity and/or prenatal developmental data, (2) in 98.5% (132/134) of the repro screening studies, the number of animals (10/sex/group) was appropriate according to the guideline for detecting DART findings and resulted in sufficient numbers of pregnant females and litters for a thorough evaluation of developmental and reproductive endpoints, and (3) in 3% (4 out of 134) of the studies the repro screening study failed to provide definitive results using the current study design.

4.4.1. Type of DART findings

Our data showed that out of 27 DART positive studies, the findings in 7 studies would also be seen in a 28-day toxicity study and the findings in 10 studies (or 7 when excluding the 3 studies with undetermined postnatal deaths) would be noticed during a prenatal developmental study. From this can be concluded that up to 48% of the cases, or at least 37% when excluding the studies with undetermined postnatal death, are not reflected in the 28-day and/or prenatal developmental toxicity study. Therefore, it would not be recommended to waive a repro screening study when negative results for a 28-day and/or prenatal developmental toxicity study are available, as suggested by Aulmann [Aulmann 2012]. Despite this, following the order of the guideline testing order, a prenatal developmental toxicity study is usually started after the point at which a repro screening study is required.

4.4.2. Sufficient number of animals per group

In all 27 repro screening studies that showed developmental and/or reproductive effects, the study design resulted in clear and conclusive treatment-related DART effects as an average of 79% of the animals or litters were affected. This shows that, when dose levels are correctly chosen, the compounds that cause developmental and reproductive toxicity show these effects in the majority of animals and/or litters. Additionally, it was investigated if 10 females/sex/group would result in a sufficient number of pregnant females (and litters) per group. Out of the 134 screening studies, there was only one case (3/10 non-pregnant females at the high dose group) in which a slightly higher number of animals per group would have been helpful for interpretation of the data. For all other studies, 10 animals/sex/group resulted in clear and definite results. In regard of executing this study as CRO, knowing 10 animals/sex/group is sufficient for the robust study design of the repro screening study, as this evaluation suggests, it would be warranted to evaluate the number of animals used in a 2-generation study. This study is even more powerful in its study design as it includes a longer treatment period, more parameters and two generations.

4.4.3. Limitations and recommendations

Ambiguous study results

Out of the 134 repro screening studies, the findings observed in 4 studies (i.e. 3%) were considered ambiguous. For 2 of these studies, the equivocal findings were considered ambiguous because of the relatively low number of affected animals (i.e. 2/9 and 3/10). In the 2 remaining studies, sex ratio was the only parameter affected. As this is a standalone parameter comprising just one value, the laboratory should have ample historical control data for a correct interpretation. Both studies showed a sex ratio that was noted only once in the historical control data at the ultimate limit (see Figure 4.6), and therefore a possible relationship to treatment could not be disregarded. It should be noted that a change in sex ratio may be indicative of a change in anogenital distance leading to a 'misclassification' of the sex, rather than a direct effect on sex ratio. Therefore, including determination of anogenital distance and internal sex examination as standard parameters to the OECD 421 and 422 guidelines would add predictive value. These parameters are very easy to assess, not expensive, do not require additional animals, and will significantly improve interpretation of sex ratio changes. For those 2 studies with only a possible effect on sex ratio, these additional endpoints would have been essential for a definitive conclusion on whether the change was related to the test compound.

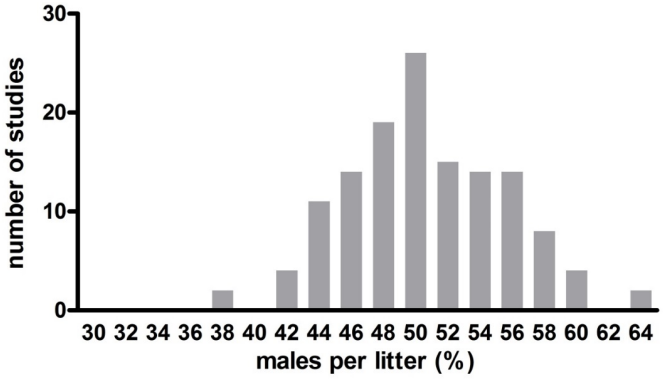


Figure 4.6. Historical control data sex ratio (134 repro screening studies).

Prolongation of postnatal period

Prolongation of the postnatal period could not be assessed as limitation by the use of the repro screening studies since none included a prolonged postnatal period. Therefore, to answer this important question, this evaluation was performed using one- and two-generation studies. Data were used from 14 two-generation and 5 one-generation studies (performed with different test compounds); these GLP studies were performed at the same testing facility during 2000 to 2009. Out of the 8 generation studies with DART findings, 5 of them showed affected pup body weights only, which all occurred after Day 4 of lactation. This was a strong indicator that prolongation of the postnatal period, at least until weaning, would be crucial for detecting of late postnatal effects

(i.e. effects potentially induced or manifested after postnatal day 4). Translating these results to repro screening studies, it would be very simple to extend the postnatal period in these repro screening studies until Day 21 of lactation, which would provide substantially more data on postnatal development but will not increase the number of experimental animals used. Interestingly, solving the previous two limitations by adding these extra parameters would also give insight on endocrine disruption. Anogenital distance and internal sex examination would provide conclusive information on sex ratio, and prolongation of the postnatal period would enable inclusion of nipple retention assessment. This could be of great value in the weight-of-evidence approach among others proposed by Testai et al. [Testai 2013]. They state that if a chemical meets the criteria for an endocrine disrupter in experimental animals, the relevance of observed effects to the human then needs to be addressed. Addition of these parameters could already provide a first indication if a chemical has endocrine disrupting potential. Practically, these additions are very straightforward, do not require extra animals, will significantly improve interpretation of sex ratio changes, will provide substantially more information on postnatal development and could give an indication for possible endocrine disruption.

Addition of recovery phase

Addition of a recovery phase to a combined 28-day/repro screening study should be done with consideration. The method described in the guideline (addition of non-mated recovery animals) can be inadequate in the case of findings influenced by pregnancy or lactation (as data on lactating main females will be compared with data of non-mated recovery females); during pregnancy and lactation physiological changes occur, kinetic data can be affected, and additional physiological stress occurs. This can be solved by adding extra non-mated main females. However, this would increase the total number of animals from 80 to 120, which is a major disadvantage of this approach. The third approach is to add recovery males only; though no recovery can be assessed in the case female specific findings are seen. Overall, there are three approaches to add a recovery phase. The choice of which depends on available data of the compound or related substances (e.g. possibility of sex specific findings or findings influenced by pregnancy or lactation).

4.4.4. Limitations of the current retrospective evaluation

The main limitation for performing a retrospective evaluation on repro screening studies at this moment is that data for comparing repro screening studies with generation studies, using the same test compound, is scarce. As in several years (after the 2018 REACH deadline), more data on this will become available and future comparisons can be made on the added value of a generation study compared to a screening study and to evaluate the possible shortcomings of repro screening studies. Prior to that, it is of essential importance to characterize the merits and pitfalls of the repro screening study now, on itself, in order to understand better, and improve, the repro screening study, instead of 'waiting' if we did right. An attempt to compare repro screening studies with generation studies was already done in 2003 by Reuter et al. [Reuter 2003]; however, they recognized that this was an impossible task because of lack of a sufficient database at that time. Their database consisted of 57 chemicals of which only 5 chemicals showed reproductive

toxicity. All 5 substances showed evidence of potential fertility impairment and unequivocal signs of developmental toxicity. These results confirmed the outcome from the preceding full-scale reproductive/prenatal developmental toxicity studies, and gave already a first indication that repro screening studies are more valuable than generally thought.

4.5. CONCLUSION

In this retrospective evaluation of 134 repro screening studies, their effectiveness was assessed. A substantial portion of the DART findings, for up to half of the studies, might be missed in other types of studies. Therefore, it should be stated that waiving a reproduction toxicity study based on the availability of a 28-day and/or prenatal developmental toxicity study could lead to misinterpretation of the test compound safety. Overall, repro screening studies are effective in providing unique data; especially considering the limited number of animals utilized, reflecting the robustness of the study design. Based on this extensive database, it was concluded that compounds causing developmental and reproductive toxicity show these effects in the majority of animals and/or litters. A simple addition of measurements, as anogenital distance, internal sex determination and nipple retention plus extending the postnatal period until weaning will add predictive value to improve the study efficacy and preventing ambiguous study results.

Acknowledgements
The authors would like to thank the involved personnel of WIL Research Europe B.V. for their highly qualified and dedicated work.

Chapter 5

Update of OECD DART guidelines with endocrine disruptor relevant endpoints: practical considerations

Manon Beekhuijzen, Francois van Otterdijk, Willemien Wieland, Miranda van Tuyl, Robert Pels Rijcken, Birgit Peter, Harry Emmen

WIL Research Europe B.V., 's-Hertogenbosch, The Netherlands.

Published in: Reproductive Toxicology. 2016 Sep;64:64-71. doi: 10.1016/j.reprotox.2016.04.002. Epub 2016 Apr 7.

ABSTRACT

In 1998, the OECD initiated a high-priority project aimed at revising existing test guidelines and developing new test guidelines for screening of potential endocrine disruptors. In 2011, OECD 443 was adopted, and in 2015 OECD 421 and OECD 422 were updated with endocrine disruptor relevant endpoints. A feasibility study for the enhancement of OECD 414 with endocrine disruptor relevant endpoints is currently ongoing. The addition of these endpoints is considered crucial for gaining more information on endocrine disruptor potency of tested chemicals, however it should be noted that these additions have a major impact on the study designs and give rise to several practical challenges. The aim of this review is to discuss important aspects of these challenging study designs and to share our knowledge on their implementation in our laboratory. Together, this review can be used as guidance for other laboratories, study monitors and registration officers.

Keywords
Endocrine disruption, OECD 421, OECD 422, OECD 443, repro screening, EOGRTS, extended one-generation reproductive toxicity study, regulatory

5.1. INTRODUCTION

Almost 20 years ago, an important endeavor was started by the OECD to update existing test guidelines and to create new guidelines for the screening and testing of potential endocrine disruptors [OECD report 1998]. In this context, parameters suitable for the detection of endocrine disrupting activity of test chemicals were added to OECD 407 (Repeated Dose 28-Day Oral Toxicity Study in Rodents) in 2008 [OECD 407]. More recently, OECD 443 (Extended One-Generation Reproductive Toxicity Study; EOGRTS) was implemented as new guideline in 2011 to replace OECD 416 (Two generation reproduction toxicity study) [OECD 416]. In 2015, OECD 421 (Reproduction/Developmental Toxicity Screening Test) and OECD 422 (Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test) were revised to include endocrine disruptor relevant endpoints [OECD 421, OECD 422, OECD 443]. A feasibility study for the enhancement of OECD 414 (Prenatal Developmental Toxicity Study) [OECD 414] with these parameters was launched in April 2015, and is now in progress.

Prior to the revision of OECD 407, there was an extensive international program to test for the relevance and practicability of the additional parameters, the performance of these parameters for chemicals with (anti-)oestrogenic, (anti-)androgenic, and (anti-)thyroid activity, the intra- and interlaboratory reproducibility, and the interference of the new parameters with those parameters already required by the previous version of the OECD 407 guideline. The large amount of data obtained has been compiled and evaluated in detail [OECD report 2006], therefore no further elaboration will be given on OECD 407, as the current manuscript will focus on Developmental and Reproductive Toxicity (DART) studies.

The selection of endocrine disruptor relevant endpoints included in the updated OECD 421 and 422 guidelines (and also in OECD 443) were based on a feasibility study addressing scientific and technical questions related to their inclusion, as well as possible adaptations of the test design needed for their inclusion [OECD No. 217, 2015]. Table 5.1 provides an overview of the endocrine disruptor relevant endpoints added to OECD 421, 422 and 443. One notable point is that the EPA guidelines for the Reproduction/Developmental Toxicity Screening Test and Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test (OPPTS 870.3550 and 870.3650) have not yet been updated with these parameters [EPA 870.3550, EPA 870.3650].

The addition of these endpoints is considered crucial for gaining more information on the endocrine disruptor potencies of tested chemicals. However, it should also be noted that these additions have a major impact on the study designs and give rise to several practical challenges, especially as there are currently a growing number of OECD 421 and 422 studies being performed for REACh. These recent additions have a huge impact on laboratory capacities as the studies are more labor-intensive and now take approximately 11 weeks instead of 7 weeks for the in-life phase.

Table 5.1. Overview of endocrine disruptor relevant endpoints added to OECD 421, 422 and 443.

OECD 421 and 422	**OECD 443**
Vaginal smears are examined daily for a period of 2 weeks prior to the onset of dosing. Only animals displaying a 4- to 5-day cycle continue in the study.	---
Oestrous cycle determination two weeks pre-mating until evidence of mating.	
The anogenital distance (AGD) of each pup is measured on the same postnatal days (between PND 0 and PND 4).	
The number of nipples/areolae in male pups is counted on PND 12 or 13.	
Blood samples are taken from: • 2 pups (one pooled sample) per litter on PND 4 for possible future analysis • 1 pup/sex/litter on PND 13 for T4 • all adult males at termination for T4 • all dams at termination for possible future analysis Further assessment of blood samples from the dams and PND 4 pups as well as of other thyroid hormones (TSH) are done if considered relevant.	Blood samples are taken from: • 2 pups (one pooled sample) per litter on PND 4 (if considered necessary) for T4. • 1 pup/sex/litter on PND 22 (F_1-pups not selected for cohorts) for T4 and TSH. • 10 F_0-males and females per dose group at termination for T4 and TSH. • 10 Cohort 1A males and females per dose group at termination for T4 and TSH.
Thyroid weight and histopathological examination not required in OECD 421; this is required in OECD 422.	Thyroid weight and histopathological examination required.
---	3 additional endocrine disruptor relevant endpoints are included in the OECD 443, namely vaginal patency, balano-preputial separation, and day of first estrus.

Implementing the new OECD 443 study and the changes in OECD 421 and 422 studies at our laboratory has revealed important practical aspects, knowledge of which can be invaluable for other laboratories. To be in accordance with the 3R's, the authors of this manuscript have therefore decided to share a review of the best practices, insights and guidance for the most important aspects of these complex study designs. Such information would be highly valuable for study monitors and registration officers.

This manuscript focuses on endocrine disruptor relevant endpoints, and discusses the following topics: legislation, practical aspects of the recently included endocrine disruptor relevant endpoints to OECD 421, 422 and 443, details on this for the OECD 414 guideline, and other important considerations. All other important practical considerations relevant to OECD 443 have recently been published [Beekhuijzen 2016].

5.2. LEGISLATION

In 2012, the OECD 421 and 422 test guidelines were updated by the Joint Meeting of the OECD council, which agreed on the status of deleted and former versions of Test Guidelines with regards to the Mutual Acceptance of Data (MAD). An 18-month transitional period between the Council Decision and the effective date of deletion was agreed, after which no new test using the deleted or the former version of a Test Guideline can be initiated. Thus if a test is initiated during the transition period (i.e. 28 July 2015 to 28 January 2017), the previous version of an OECD test guideline will be applied, and the test results would be acceptable under the principles of MAD. Evaluating authorities might however request additional data upon evaluation.

With regard to data evaluation, it is important to note that, to date, there is no globally accepted strategy to identify endocrine disruptors. As a consequence, the data requirements for addressing potential endocrine disruption in humans differs depending on the type of substance being registered and the region of interest. To illustrate this, the requirements for industrial chemicals, biocides and agrochemicals in the European Union (EU) vs. United States (U.S.) are highlighted below (a full overview of global requirements is outside the scope of this manuscript).

In the EU, under REACh law, every industrial chemical registration at 10 tonnes/annum or more should include test data to address potential effects on development and reproduction. The minimum data requirement for Annex VIII (10-100 t/a) consists of a repro screening study (OECD 421/422), at higher tonnages the prenatal developmental toxicity study [OECD 414] and the EOGRTS [OECD 443] may also be required. The current dossier format does not include a separate endpoint to address potential endocrine disrupting activity of a substance, but this endpoint should be assessed within other sections (e.g. Toxicity to Reproduction). In the absence of well-defined criteria for endocrine disruption, a well-defined testing strategy is lacking, although the OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters [OECD No. 150, 2012] can be used as guidance for testing. Upon evaluation by the European Chemicals Agency (ECHA), endocrine disrupting activity can however be a factor that determines if a substance is regarded as a Substance of Very High Concern (SVHC), which can trigger a ban from the market.

The U.S., unlike the EU, does not have fixed data requirements for industrial chemicals. New chemicals can be added to the Toxic Substances Control Act (TSCA) Chemical Substance Inventory through the submission of a Premanufacture Notice (PMN), in which all currently available data should be included. After submission, the Environmental Protection Agency (EPA) evaluates the data via the application of several tools including, but not limited to, structure activity relationships, data from analogues, predicted mechanisms of toxicity, and expert judgment. As a follow-up, the EPA can request further information, however it should be noted that almost 90 percent of Premanufacture Notices submitted to the EPA complete the review process without being restricted or regulated in any way [EPA source]. As a result, for a number of substances used in U.S. commerce (irrespective of the tonnage), potential effects on reproduction and/or development might

not have been addressed. The EPA furthermore registers (and licenses) all pesticides distributed or sold in the United States, as regulated in majority under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and its subsequent amendments. Before EPA may register a pesticide under FIFRA, the applicant must show, among other things, that using the pesticide according to specifications "will not generally cause unreasonable adverse effects on the environment'', which includes any unreasonable risk to man. To this end, several Health Effects Test Guidelines are available.

In Europe, information requirements for biocidal products are set in the Biocidal Products Regulation, Regulation (EU) 528/2012) (BPR), which states that if there are indications that a substance possesses endocrine disrupting potential, this aspect has to be further elucidated by test work.

For the registration of agrochemicals in the U.S., the EPA has the Endocrine Disruptor Screening Program (EDSP) in place, which applies a tiered approach for screening agrochemicals, i.e. a Tier 1 screening battery and Tier 2 testing approach focused on estrogen, androgen and thyroid pathways in humans and wildlife. Based on risk assessment taking into account all data, the EPA can make risk management decisions regulating the chemical.

In the EU, agrochemicals are evaluated in a multi-step process: one EU country is appointed as Rapporteur Member State (RMS) and prepares a draft assessment report, based on which a conclusion is drawn by the European Food Safety Authority (EFSA) and the Member States. In a final step, the Committee for Food Chain and Animal Health votes on approval or non-approval. Data requirements for agrochemicals include assessment of several endpoints related to potential endocrine activity. EFSA applies the definition for an endocrine disruptor as proposed by the World Health Organisation [IPCS 2002], and addressing this endpoint is part of the risk assessment. Every chemical that is regarded as having endocrine disrupting properties that may be harmful to humans or non-target organisms, is in principal banned from the EU market.

5.3. ENDOCRINE DISRUPTOR RELEVANT ENDPOINTS ADDED TO OECD 421, 422 AND 443

This chapter will describe essential details on the conduct of the six endocrine disruptor relevant endpoints: oestrous cycle, anogenital distance, nipple retention, thyroid hormones, necropsy and histopathology, and three parameters added solely to the OECD 443 (i.e. vaginal patency, balanopreputial separation and day of first vaginal oestrus).

5.3.1. Oestrous cycle

Vaginal smears should be monitored daily (i.e. collected and examined for oestrous cycle stage) during two weeks of the treatment period until evidence of mating for OECD 421, 422 and 443 studies.

Additionally, the updated OECD 421 and 422 guidelines recommend to start each group with at least 10 males and 12-13 females (the former guidelines used 10 animals/sex/group). Females will be evaluated pre-exposure for oestrous cyclicity and animals that fail to exhibit at least 2 regular 4-5 day oestrous cycles during this 2-week pretreatment period will not be included in the study. Therefore, extra females are recommended in order to guarantee that 10 females can be selected prior to start of treatment in each group. This approach is not required for the OECD 443, probably due to the larger number of animals used in the latter study.

In recent months, our laboratory has performed 11 updated OECD 421/422 studies in which only 6 out of 528 females (Wistar Han) showed an irregular estrous cycle during the 2-week pretreatment period. From this it can be concluded that pre-selection is of limited value when balancing the additional number of animals used versus improvement of the test system. In the coming period, this should be evaluated thoroughly and, if shown to be unnecessary, adjusted in the guidelines.
In addition, selection of females with regular estrous cycles shortly before start of the treatment period makes the OECD 421/422 studies more labor intensive. This is because oestrous cyclicity should be determined without delay and a few days before start of treatment the decision for allocation should be made and confirmed by the study director. Starting a study on a Monday or Tuesday is therefore less practical as it would mean that this additional work needs to be done on the weekend. From the 11 updated OECD 421/422 studies performed in our laboratory, in most cases (7x) a decision could be made 1 day before start of treatment (i.e. on Day 14). For two studies, this decision could be made on Day 13 and in the remaining two studies on Day 15 (i.e. on the first day of treatment).

Apart from using additional animals, this examination will also lead to extra discomfort for all female animals. It is therefore important to use the least invasive method for collection of vaginal cells. At WIL Research, we use vaginal lavage instead of swaps. This method is known to result in reduced inductions of pseudopregnancy. We have not seen any pseudopregnancies in the 11 updated OECD 421/422 studies performed in our laboratories. At WIL Research, we use the following procedure for collecting vaginal cells:

- The rat is removed from its cage and held in a supine position with one hand around the thoracic area while also restraining the tail. The vaginal area is exposed to the individual who is preparing to lavage the animal. An alternative restraining method for obtaining a lavage sample is as follows: After checking the identification of the animal, grasp the animal at the base of the tail and place the animal in a vertical position, head facing down, with the hind paws resting on the ventral area of the individual lavaging the animal. The vaginal area will be exposed using this technique.
- Draw up approximately 0.1-0.2 mL of deionized water in a new disposable eyedropper or small pipette and gently insert the eyedropper or pipette into the vagina, no deeper than 0.5 cm.

- Aspirate the deionized water in and out of the vagina 2-3 times. Excessive aspiration can cause disruption of the cell arrangement or damage the existing cells in the lavage sample.
- Carefully drop the vaginal rinse from the eyedropper or pipette on a labeled microscope slide and drag the vaginal rinse drop approximately 2-3 cm in a straight line. Dry, fix and stain the lavage slides with Giemsa stain.
- Use a new eyedropper or pipette for each lavage sample.

Important to note that for the first few studies, we allocated the selected female animals (i.e. with regular oestrous cycle) to their definitive dose group 1 or 2 days before start of treatment. This meant that allocation (weighing of the animals), identification (tattooing) and replacement of their cage in the rack/room was done usually on the same day and shortly before start of treatment. This together with daily vaginal lavage and start of treatment caused too much stress to the female animals. As a result, we changed our methods and are now allocating the female animals already before the start of the 2-weeks pretreatment period, and if present, replacing female rats with irregular cycles 1-2 days before start of treatment.

5.3.2. Anogenital distance (AGD)

All three guidelines mention that the AGD of each pup should be measured on the same postnatal day between PND 0 and PND 4. Pup body weight should be determined on the day the AGD is measured. The AGD should also be normalized to pup size, preferably the cube root of body weight. AGD is sometimes standardized to crown-rump-length.

At WIL Research, the anogenital distance for each neonate is measured on PND 1 using a caliper with a digital read out, from the caudal margin of the anus to the caudal margin of the genital tubercle. Measurements are commonly recorded to the nearest 0.01 millimeter. The tail of the pup is held gently between the thumb and ring/little finger and the torso is restrained between the first two fingers of the left hand. It is extremely important not to stretch the AG region. See Table 5.2 for historical control data.

Table 5.2. Historical control data of AGD (mm) collected at WIL Research Europe, The Netherlands (Wistar Han rats; 11 studies).

	Mean	SD	N litters	P5 percentile	P95 percentile	Normalized*
Male pups	2.55	0.353	98	2.18	3.11	1.55
Female pups	0.97	0.229	96	0.67	1.40	0.61

* AGD normalized to cube root of body weight

5.3.3. Nipple retention

For all three guidelines, the number of nipples/areolae should be counted for male pups on PND 12 or 13 as recommended in OECD guidance document (GD) 151 [OECD GD 151, 2013].
GD 151 mentions that because hair growth makes it difficult or impossible to see the areolae, it

is important to establish the correct time for the assessment. The presence of nipples/areolae in male pups should be measured when they are obvious (i.e. as they appear in the female litter mates), ideally on PND 12 or 13 (time may vary with strain). As far as possible, all pups should be evaluated on the same PND as there can be marked differences as maturation progresses. A quantitative count in male pups is also recommended, as a qualitative assessment only (presence/absence) of nipples/areolae may be rather insensitive particularly when control incidence is high. At WIL Research Europe, we perform a quantitative count on PND 13 for Wistar Han rats.

5.3.4. Thyroid hormones

Thyroid hormones are extremely important for fetal and neonatal growth and development. There are several differences for the measurement of thyroid hormone levels between the OECD 421/422 versus OECD 443 guidelines.

For adult animals, blood samples are taken from 10 animals/sex/group at termination. During the EOGRTS (OECD 443), this is done for the F_0-animals and Cohort 1A animals. For the OECD 422 and 443, this is combined with blood sampling for hematology and clinical biochemistry, and therefore the animals will be fasted overnight. In theory, overnight fasting would not be needed for an OECD 421 study by default. However, as it is important to collect historical control data under identical circumstances, we felt compelled to also fast the animals from the OECD 421 overnight before blood sampling for thyroid hormone measurement. This change in the study design of the OECD 421 causes extra discomfort for all animals.

For the EOGRTS, these blood samples will be used for determination of T4 and TSH concentrations. This is however different for OECD 421/422 studies as only T4 determination is required for male animals. Measurement of T4 in blood samples from the dams is only done if relevant, and other hormones may also be determined if relevant. However, there are no criteria defined that prompt the analysis of additional hormones.

For pups, there are two occasions for blood collection for (possible) thyroid hormone level determinations. On PND 4, this is required for all three guidelines. The second time point is at termination which is in the OECD 421/422 on PND 13 and in the OECD 443 on PND 22.

At PND 4, one blood sample is collected from two pups per litter and pooled into one sample (0,3 mL blood is collected at WIL Research Europe by decapitation) for possible determination of T4 levels. OECD 421/422 states that T4 assessment in these pups should be performed if relevant. OECD 443 mentions that for the remaining pups at PND 4, measuring serum thyroid hormone (T4) concentrations should also be considered. However none of the three guidelines mention when these measurements should be included.

At termination, blood (0.9 mL at WIL Research Europe by aorta puncture) is collected for T4 and TSH analysis from PND 13 pups in OECD 421/422 studies and from weanlings on PND 22 (F_1-pups

not selected for cohorts) in the OECD 443. In the EOGRTS, T4 and TSH measurements are both required according to OECD 443. However, the OECD 421/422 guidelines only require T4 measurement in first instance. Again, these guidelines do not state when TSH analysis should be performed. The guidance document for the EOGRTS (GD 151) states that "statistically or biologically significant changes in either one or both hormones between treated and control groups can be evaluated in conjunction with any changes in thyroid gland weight and histopathology, as well as neurological or other developmental adversities used for risk assessment" [OECD GD 151, 2013]. For OECD 421/422 studies, the procedures described in Table 5.3 are implemented at WIL Research Europe.

Table 5.3. Decision tree for additional thyroid hormone measurements.

Generation	Time point of terminal blood sampling	Information available	If statistically significant or relevant effect	Then measure
F_1-generation (2 pups/litter, pooled)	PND 4	---	T4 PND 13 pups	T4
F_1-generation (2 pups/litter)	PND 13	T4	T4	TSH
F_0-males	After at least 28 days of treatment	T4 Thyroid weight Histopathology thyroid	T4 Histopathology thyroid	TSH
F_0-females	PND 14	Thyroid weight Histopathology thyroid	T4 F_0-males Histopathology thyroid	T4 and TSH

In order to make well informed decisions, we also decided to perform histopathological examinations of the thyroids in the OECD 421 studies, even though this is not a guideline requirement (it mentions "may be examined when necessary"). Without histopathological analysis, the authors felt that the decision to measure these hormone levels would be based on insufficient grounds, as it is known that histopathological examination of the thyroid is usually more sensitive than thyroid weight and hormone levels. The validation report of OECD 407 [OECD report 2006] states that "thyroid histopathology was consistently the most reliable and most sensitive endpoint for the detection of thyroid modulation. Thyroid weight was reliable, but was somewhat less sensitive when compared to thyroid histopathology. Circulating thyroid hormone levels (T3, T4, and TSH) were not always reliable and sensitive, but the standard operating procedures for blood sampling and for thyroid hormone analyses were not standardized to reduce stress induced variability and to reduce analytical variability, respectively. Circulating T4 levels were the most promising of the three thyroid hormonal values".

Selection of PND 4 pups for blood sampling is described in detail in the OECD 421/422 guidelines, but not in the OECD 443 guideline. OECD 421/422 states that "when culling is used, blood samples should be collected from two of the surplus pups, pooled, and used for possible determination of serum T4 levels. If there are insufficient pups in a litter to have two surplus pups, two of the pups

that are typically retained would be used. When litter size drops below the culling target (8 or 10 pups/litter), preference will be given to remove two female pups for PND 4 blood samples in order to retain more male pups for nipple retention on PND 13; however, retained female pups in each litter should not drop below 2 if this can be avoided. If litter size is not adjusted, preferably the two pups per litter selected for T4 blood sampling should be female pups to reserve male pups for nipple retention evaluations".

The guidelines are very specific on the minimum number of female pups. However, it does not specify the minimum number of male pups. In the case, you would decide to retain at least 4 male pups, Table 5.4 can be used for pup selection.

Table 5.4. Procedure for pup selection for blood sampling.

Number of pups per litter	Number of male pups per litter	Number of female pups per litter	Selected two pups for T4 blood
0	---	---	None
1	0 to 1	0 to 1	None
2	0 to 2	0 to 2	None
3	0 to 3	0 to 3	None
4	0	4	Two female pups
	1 to 4	1 to 3	None
5	0 to 1	4 to 5	Two female pups
	2 to 5	0 to 3	None
6	0 to 2	4 to 6	Two female pups
	3 to 5	1 to 3	None
	6	0	Two male pups
7	0 to 3	4 to 7	Two female pups
	4 to 5	2 to 3	None
	6 to 7	0 to 1	Two male pups
8	0 to 4	4 to 8	Two female pups
	5	3	One male and one female pup
	6 to 8	0 to 2	Two male pups
9	0 to 4	5 to 9	Two female pups
	5 to 6	3 to 4	One male and one female pup
	7 to 9	0 to 2	Two male pups
9+	0 to 9+	0 to 9+	Two pups selected for culling

However as default, at WIL Research Europe we decided not to follow this part of the guideline (after approval from our clients). Instead, because of the following reasons, blood collection for possible thyroid hormone analysis from PND 4 litters with a size of ≤ eight pups will not be performed:

- Most litters will consist of > 8 pups and these samples will thus be available,
- T4 measurement for PND 4 pups only needs to be done if relevant,
- T4 is measured for PND 13 pups by default,
- And most importantly, further reducing small litter numbers can affect interpretation of other essential developmental parameters (e.g. survival, body weight).
- In case, the test substance causes reduced litter sizes and therefore no blood samples can be collected from PND 4 pups for T4, it still allows sufficient information on thyroid hormone levels as this will be determined from PND 13 pups. Moreover, especially in this case it is important to keep the litter sizes unchanged throughout development to allow a thorough evaluation of developmental toxicity.

This adjustment will in our opinion lead to more complete data, without having impact on the interpretation or validity of the data. At the time of submission of this manuscript, a proposal for adjustment of the guidelines was issued at the OECD.

For a number of measurements in serum and plasma, most notably for glucose, overnight fasting would be preferable. The OECD 421/422 guidelines and GD 151 mention that if overnight fasting is adopted, the dams should be retained for an additional day after the pups are removed. The reason for this is that overnight fasting may disturb lactation and nursing behavior. However, the stress caused by removing the pups from their mother on PND 13 for OECD 421/422 studies should not be underestimated. Additional research is needed to determine the least stressful approach for the dam and pups.

An important consideration when collecting blood samples for thyroid hormone measurement, is that the variation in the time of sacrifice should be kept to a minimum within the study and between studies (when comparing historical control data), because of diurnal variation of hormone concentrations. This can be very challenging as the dates of sampling will be defined by the date of delivery of the litters, and this cannot be planned in advance.
See Table 5.5 for historical control data.

Table 5.5. Historical control data of total T4 (µg/dL) collected at WIL Research Europe, The Netherlands (Wistar Han rats; 11 studies).

	Mean	SD	N	P5 percentile	P95 percentile
Adult males	4.68	1.108	107	2.92	6.75
PND 13 males	6.26	1.602	95	3.52	8.48
PND 13 females	5.80	1.411	92	3.68	8.14

5.3.5. End of in-life evaluations at necropsy

Table 5.6 (see page 100) provides an overview of the reproductive organs and accessory sex organs that need to be collected, weighed and/or examined histopathologically according to OECD 421, 422

and 443.
From this table it is clear that there are major differences between the three guidelines:

1) According to OECD 422, histopathological examination is only needed for five animals/sex/group compared to all animals/group in the other two guidelines. As default, at WIL Research Europe we also perform histopathological examination of the reproductive organs of all males that failed to sire and all females that failed to deliver healthy pups.
2) According to OECD 421, histopathological examination is only needed for three organs (testes, epididymides and ovaries) as the remaining information is already available from a repeated dose study (28-day and/or 90-day). However, it should be noted that where these examinations were performed a long time ago or where the quality is not up to current standards, it would be worthwhile to examine all reproductive/accessory sex organs. By default, at WIL Research Europe we perform histopathological examination of the reproductive organs of all males that failed to sire and all females that failed to deliver healthy pups. In addition, OECD 421 does not include histopathological examination of the thyroids as default; the guideline mentions "the other preserved organs including thyroid may be examined when necessary". We recommend including this as default at least for the adult animals in order to make a thorough decision as to whether additional thyroid hormone analyses are needed, as thyroid pathology is an important marker for endocrine disruption.
3) OECD 421/422 do not clearly state that weight determination of the thyroids is required, as the thyroids are not included in the list of organ weights. But there is a sentence in another paragraph (#49 in OECD 421 and #64 in OECD 422) that says that "thyroid weight could be determined after fixation". It is unclear from this statement as to whether they mean that the weights could be determined or that the weights should be determined but could be done either before or after fixation. We would always recommend the inclusion of this measurement.
4) Strikingly, only OECD 443 states the requirement of histopathological examination of the mammary glands in both males and females. This is not mentioned in OECD 421, while OECD 422 states that the male mammary gland may give a valuable indication of endocrine-related effects, but does not mention it for preservation or histopathological examination. It also states that "changes in male mammary glands have not been sufficiently documented but this parameter may be very sensitive to substances with estrogenic action". As default, at WIL Research Europe we also collect the female mammary glands in OECD 421/422 studies, so that in the case of a female with total litter loss, the reason for this loss can be further investigated.
5) OECD 421/422 states that "the testes and epididymides (as well as levator ani plus bulbocavernosus muscle complex, Cowper's glands and glans penis) of all adult males should be weighed". It is unclear as to why these organs are mentioned in brackets. However, after inquiring at the OECD working group, we found out that this apparently means that these three organs should only be weighed and not be preserved for histopathological examination.

OECD 421 and 422 mention that on PND 13 the thyroid from 1 male and 1 female pup per litter should be preserved. Again, the guidelines are not very specific on organ weight determination (i.e. the thyroid weight could be determined after fixation) or histopathology (i.e. the thyroid gland from pups may be examined when necessary).

Table 5.6. Overview of end of in-life examinations at necropsy for OECD 421, 422 and 443 studies.

	OECD 421			OECD 422			OECD 443		
Cowper's glands	P	W	--	P	W	--	--	--	--
Epididymides	P	W	HA	P	W	HS	P	W	HA
Glans penis	P	W	--	P	W	--	--	--	--
Levator ani plus bulbcavernosus muscle complex (LABC)	P	W	--	P	W	--	--	--	--
Mammary gland (males and females)	--	--	--	--	--	--	P	--	HA
Ovaries	P	--	HA	P	--	HS	P	W	HA
Prostate	P	--	--	P	--	HS	P	W	HA
Seminal vesicles including coagulating glands	P	--	--	P	--	HS	P	W	HA
Testes	P	W	HA	P	W	HS	P	W	HA
Thyroid	P	--	--	P	--	HS	P	W	HA
Uterus (including cervix)	P	--	--	P	--	HS	P	W	HA
Vagina	--	--	--	--	--	--	P	--	HA

P = preservation; W = weighing; HA = histopathological examination all animals (Groups 1 and 4 in first instance); HS = histopathological examination from selected 5 animals/sex/group (Groups 1 and 4 in first instance)

5.3.6. Vaginal patency, balano-preputial separation, and day of first vaginal oestrus

There are three endocrine disruptor relevant endpoints that are only included in the OECD 443. These are 1) vaginal patency, 2) balano-preputial separation, and 3) day of first vaginal oestrus. These are not integrated in OECD 421/422 studies, as the offspring are terminated on Day 4 or 13 post-partum.

All selected F_1-animals are evaluated daily for balano-preputial separation or vaginal patency for male or female animals, respectively. These evaluations commence before the expected day for attainment of these events to detect if sexual maturation occurs early. At WIL Research Europe, we start with these examinations on PND 25 for female Wistar Han rats and on PND 35 for male Wistar Han rats. Any abnormalities of genital organs, such as persistent vaginal thread, hypospadia or cleft penis, should be recorded. In addition, vaginal smears should be examined daily for all F_1-females in cohort 1A after the onset of vaginal patency, until the first cornified smear (i.e. first vaginal oestrus) is recorded. This is in order to determine the time interval between these two events. More detail and background information is reported in OECD GD 151 [OECD GD 151, 2013].

For all three parameters, it is very important to evaluate these data in combination with the body weights at attainment. Mothers exposed to high doses of a test material often have offspring with lower body weights. This will affect offspring maturation and hence these endpoints.

5.4. OECD 414

A feasibility study is currently ongoing for enhancement of OECD 414 (Prenatal Developmental Toxicity Study) with endocrine disruptor relevant endpoints. As was done for the update of OECD 421 and 422, a preliminary analysis of studies performed in a research context was recently conducted. This initial step enables identification of the most promising endpoints to be included in the revised guideline. Data from studies performed in Denmark are presently being collected, and the objective of the call for data was to allow the analysis to be based on a broader dataset. This information is going to be used by the lead (Denmark), who will perform a statistical analysis that would aim at evaluating which new endpoints have relevance and good reliability for the identification of endocrine active chemicals. Examples of these new endpoints are testosterone levels and anogenital distance in male fetuses. It is expected that this updated guideline will come into force in 2017.

5.5. IMPORTANT CONSIDERATIONS

This chapter will describe four additional considerations that relate to these updated/new guidelines: additional animal discomfort, internal sex determination, type of diet and bedding, and pup body weight determinations.

5.5.1. Additional animal usage and discomfort

It should be acknowledged that even though the additions to the OECD 421 and 422 guidelines are crucial for gaining more information on endocrine disruptor potencies of the tested chemicals, this update however causes additional animal usage and animal discomfort:

1. The updated guidelines recommend starting each group with at least 12-13 females per group compared to 10 females per group, as stated in the former guidelines. Females will be evaluated pre-exposure for oestrous cyclicity and animals that fail to exhibit typical 4-5 day cycles during this 2-week pretreatment period will not be included in the study. Therefore, extra females are recommended in order to yield 10 females per group.
2. Oestrous cycle determination is performed for at least 4 weeks
3. If overnight fasting is adopted, the dams should be retained for an additional day after the pups are removed. However, the stress caused by removing the pups from their mother on PND 13 for OECD 421/422 studies should not be underestimated. Additional research is required to determine the least stressful approach.

4. Overnight fasting would not be needed for an OECD 421 study by default. However, as it is important to collect historical control data under identical circumstances, we felt compelled to also fast the animals from the OECD 421 overnight before blood sampling for thyroid hormone measurement.

In the coming years, there needs to be further evaluation(based on additional research and experience with these new study designs) in order to determine whether or not changes to the guidelines are essential and worth the additional animal usage and discomfort.

5.5.2. Internal sex determination

The guidelines do not explicitly state whether at necropsy of the pups, that internal sex determination should be performed. Such an assessment is very simple to include and will significantly improve interpretation of sex ratio changes (i.e. differences in external and internal sex determination after exposure to strong acting endocrine disruptor compounds) [Beekhuijzen 2014], which is especially important for the evaluation of endocrine disruption. We therefore highly recommend its inclusion.

5.5.3. Type of diet and bedding

OECD 443 provides details on the type of diet and bedding required in the context of endocrine disruption (i.e. with reduced oestrogenic substances). This requirement is not included in the OECD 421/422 guidelines. The authors would not recommend changing the type of diet only for these studies, as historical control data (based on standard laboratory diet and bedding) are crucial for the correct interpretation of the obtained results. For such studies, it is therefore recommended to request analytical certificates from the supplier for each batch of diet and bedding, and to assure that each batch contains < 325-350 µg/g total genistein equivalents (see GD 151 for details).

5.5.4. Pups body weight determinations

The OECD 443 guideline states that live pups are weighed individually on PND 0 or PND 1, and regularly thereafter, e.g. at least on PND 4, 7, 14, and 21. The OECD 421/422 guidelines state that litters should be weighed within 24 hours of parturition (PND 0 or 1) and at least on PND 4 and PND 13. As pup body weights are highly valuable as a developmental parameter, we recommend including at least the pup body weight determinations on PND 7 in OECD 421/422 studies.

5.6. CONCLUSIONS

In this review, we evaluated the impact of the endocrine disruptor relevant endpoints that are included in the EOGRTS (OECD 443), Reproduction/Developmental Toxicity Screening Test (OECD 421) and Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test (OECD 422). Furthermore, we addressed the specific considerations that should be taken into account when implementing these study designs. The shortcomings in the upda-

ted OECD 421 and 422 test guidelines were also highlighted, and recommendations to amend these were proposed. The complexity of these studies should not be underestimated. Experienced testing laboratories that have sufficient resources and historical control data for all parameters should conduct these studies. In the future, it will be important to gather and review the data from these studies in order evaluate and optimize the designs if needed, and to gain more clarity on whether or not certain parameters should be analyzed.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

The authors would like to thank their colleagues from the technical departments at WIL Research Europe, The Netherlands, for their valuable contribution and dedication to the implementation of these three study designs at our laboratory. In addition, the authors are grateful to Nima Davoodi (WIL Research Europe) for critical editorial review of the manuscript.

Chapter 6

Implementing the Extended One-Generation Reproductive Toxicity Study (EOGRTS). Important points to consider

Manon Beekhuijzen[1], Helma Barentsen[1], Edward Marsden[2], Amy Zmarowski[1], Michel Aujoulat[2], Catherine Picut[3], Eddie Sloter[4]

[1]WIL Research Europe, Den Bosch, The Netherlands.

[2]WIL Research Europe, Lyon, France.

[3]WIL Research US, Hillsborough, North Carolina, USA.

[4]WIL Research US, Ashland, Ohio, USA.

Published in: Critical Reviews in Toxicology. 2016;46(4):332-47. doi: 10.3109/10408444.2015.1137863. Epub 2016 Mar 3.

ABSTRACT

The hallmark of the EOGRTS is that based on certain criteria or triggers, selected offspring are assigned at weaning to different cohorts for further investigation of sexual maturation, reproductive organ integrity and function, neuropathological and behavioral endpoints, and/or immune function. The triggers allow for a more customizable design based directly on the data, while minimizing animal usage. Compared to the Two-Generation Reproductive Toxicity Study, the EOGRTS design increases the number, extent and duration of F_1 offspring assessments resulting in more thorough and efficient utilization of the F_1 generation while excluding the second generation of offspring unless triggered. Therefore, the EOGRTS has the potential to reduce the number of rats required by nearly 1200 animals per study. When performing the EOGRTS, the complexity of this study should not be underestimated and experienced flexible testing laboratories with sufficient resources and historical control data for all parameters are essential. The aim of this review is to discuss the important aspects of this challenging study design and to share our knowledge on the implementation of this study in our laboratories. In addition, we elaborate on the type of criteria for expansion of the study and logistical considerations. Altogether, this review can be used as guidance by other labs, study monitors and registration officers.

Keywords
OECD 443, EU B.56, Extended One-Generation Reproductive Toxicity Study (EOGRTS), 3R's, developmental neurotoxicity, developmental immunotoxicity, triggers.

6.1. INTRODUCTION

Compared to the Two-Generation Reproductive Toxicity Study (Two-Generation study) [OECD 416; EU B.35; OPPTS 870.3800], which requires F_1 mating to produce a second generation, the Extended One-Generation Reproductive Toxicity Study (EOGRTS) [OECD 443; EU B.56] design increases the number, extent and duration of F_1 offspring assessments resulting in more thorough and efficient utilization of the F_1 generation while excluding the second generation of offspring unless triggered. The EOGRTS has the potential to reduce the number of rats required by nearly 1200 animals per study (depending on the litter size of the strain). At weaning, selected F_1 offspring are assigned to specific subgroups (Cohorts 1-3; see Figure 6.1) for further investigation of sexual maturation, reproductive organ integrity and function (Cohort 1; reproductive toxicity), neuropathological and behavioural endpoints (Cohort 2; developmental neurotoxicity; DNT), and immune function (Cohort 3; developmental immunotoxicity; DIT). In addition, the need for a second generation depends on the regulatory requirements (i.e. certain specific criteria based on existing information or study results already obtained), and may be produced for instance if triggered by reproductive/developmental effects or ambiguous results. Additionally, for some compounds, separate developmental immunotoxicity and neurotoxicity (DIT and DNT, respectively) studies may no longer be needed, making the new EOGRTS design a promising alternative to multi-generational designs and an important advancement for the 3R's.

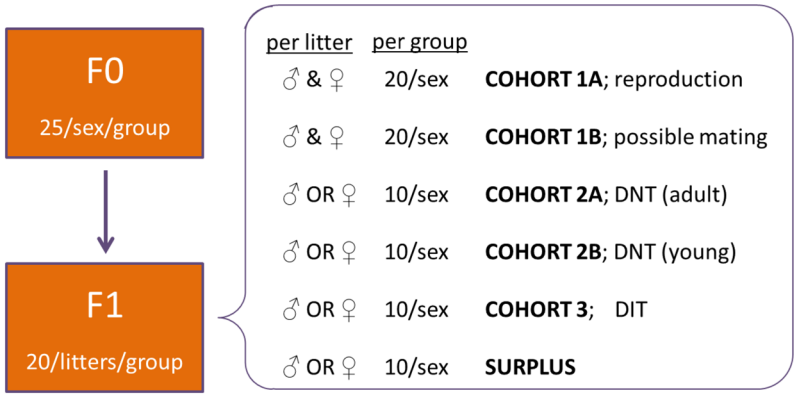


Figure 6.1. Assignment of selected F_1 offspring to the EOGRTS cohorts.

After finalization of the OECD 443 guideline, the authors were asked to explore the implementation of this study type at WIL Research and to harmonize details of the study design at the three different sites. During this challenging process important considerations in relation to guidelines/guidance documents, logistics and practical aspects were identified and procedures were implemented to overcome these. The authors felt that this wealth of information was of invaluable

use for other laboratories planning to implement the OECD 443 and to be in line with the 3R's, therefore decided to share best practices and publish the current manuscript as a guidance. Moreover, also for study monitors and registration officers this review offers insights and guidance for important aspects of this complex study design.

The manuscript discusses the following topics: legislation, triggering and waiving of the different cohorts, main components of the EOGRTS, logistics, and other important considerations.

6.2. LEGISLATION

The EOGRTS was first proposed by Cooper et al. [Cooper 2006], and the study design underwent several modifications before being adopted by the OECD as Test Guideline (TG) 443 in July 2011 [OECD 443]. In August 2014 the EU test method Regulation 440/2008 was amended by Regulation 900/2014 to include the EOGRTS as test method B.56 [EU B.56], which is equivalent to OECD 443. In February 2015, the REACH annexes VIII, IX and X were amended [EU 2015/282] and the EOGRTS is now the information requirement for reproductive toxicity under REACH instead of the Two-Generation Study [OECD 416; EU B.35; OPPTS 870.3800]. The updated annexes entered into force 13 March 2015. These mention that the EOGRTS has to be performed under REACH at a tonnage level of more than 1000 T (Annex X) always and at a tonnage level of 100-1000 T (Annex IX) "if the available repeated dose toxicity studies (e.g. 28-day or 90-day studies, OECD 421 or 422 screening studies) indicate adverse effects on reproductive organs or tissues or reveal other concerns in relation with reproductive toxicity". To our knowledge, within the legislation of other regions or even within other frameworks within Europe, such as biocides, agrochemicals and pharmaceuticals, the EOGRTS has not been incorporated yet as primary testing method, although Canada and the USA can request the EOGRTS to be performed. The EU Biocidal Products Regulation [EU 528/2012] does state in Annex II point 8.10.2 that "the extended one-generation reproductive toxicity study adopted at OECD level shall be considered as an alternative approach to the multi-generation study". EOGRTS specific legislation documents are described in Table 6.1.

With regard to classification for reproductive toxicity as established in the CLP Regulation [EC 1272/2008], sexual function and fertility, developmental toxicity of the offspring and effects on or via lactation are evaluated in the EOGRTS. And therefore, for risk assessment the EOGRTS is acceptable within all EU frameworks.

The European Commission's amendment of February 2015 [EU 2015/282] indicates the "standard information requirement in Annexes VIII, IX and X to Regulation (EC) No 1907/2006 [EC No 1907/2006] should be limited to the basic configuration of EOGRTS" (no extension and no cohorts), however "in certain specific cases, where justified" the protocol will include one or more of the additional cohorts. To implement this, it is important to develop Weight of Evidence (WoE) considerations for determining the necessity of one of the cohorts or second generation mating

within the REACH framework. The WoE should include evaluation of mode of action, structural alerts, toxicokinetics, repeat-dose results, data on analogues, in vitro data, and relevant human data of the test compound. These evaluations will be performed before the start of the study. The ECHA guidance document [ECHA R.7a, 2015] on information requirements and chemical safety assessment states that the study design should be decided before the study is started and does not recommend the use of in-study ("internal") triggers, although expansion (not reduction) of the study is allowed after the ECHA evaluation decision has been issued. Triggers for all cohorts have been included in the ECHA guidance.

Table 6.1. Overview on regulations, guidelines and guidance documents for the EOGRTS.

OECD guideline EOGRTS	OECD 443. OECD guideline for the testing of chemicals. No. 443. Extended one generation reproductive toxicity study. July 2011
EU guideline EOGRTS	EU B.56 Extended one-generation reproductive toxicity study. EU, Commission Regulation (EU) 900/2014 amending, for the purpose of its adaptation to technical progress, Regulation (EC) 440/2008 laying down test methods pursuant to Regulation (EC) 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 15 July 2014.
OECD guidance for EOGRTS	OECD GD 151. Guidance Document no. 151 supporting OECD test guideline 443 on the extended one generation reproductive toxicity test. ENV/JM/MONO(2013)10. July 2013
OECD guidance on triggers	OECD GD 117. Guidance Document no. 117 on the current implementation of internal triggers in test guideline 443 for an extended one generation reproductive toxicity study, in the United States and Canada. ENV/JM/MONO(2011)21. July 2011
ECHA guidance	ECHA Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7a: Endpoint specific guidance, Version 4.1, October 2015
REACH regulation amendment EOGRTS	EU 2015/282. Commission Regulation (EU) 2015/282 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards the Extended one-generation reproductive toxicity study, February 2015

6.3. TRIGGERS AND WAIVERS FOR COHORTS ON DNT, DIT AND SECOND GENERATION

The standard information requirement in Annexes IX and X to Regulation No 1907/2006 [EC 1907/2006] is limited to the basic configuration of the EOGRTS (Cohort 1A and 1B). Regulation 2015/282 [EU 2015/282] also states that "In certain specific cases, where justified, the registrant should be able to propose and the European Chemicals Agency (ECHA) should be able to request the performance of the F_2 generation (extension of Cohort 1B), as well as the DNT and DIT cohorts

(Cohorts 2 and 3)". Column 2 in the REACH Annexes specifies the expansion criteria to the basic configuration of the EOGRTS. In practice, performance of a cohort will be requested as soon as doubts arise.

For the extension of Cohort 1B (mating of F_1 animals) OECD 443 refers to OECD guidance document (GD) no. 117 [OECD GD 117, 2011] but REACH includes different criteria. Extension of cohort 1B to include the F_2 generation is required by ECHA if the substance has uses leading to significant exposure to consumers or professionals, and the substance displays genotoxic effects in an in vivo somatic cell mutagenicity test (resulting in Mutagen category 2 classification), or there are indications that the steady state concentration of the substance and/or metabolite(s) is only reached after extended exposure, or there are indications of one or more relevant modes of action related to endocrine disruption [ECHA R.7a, 2015]. Appendix R.7.6-2 of the ECHA guidance specifies what the above mentioned triggers might consist of.

The authors agree that a second generation (F_2) should always be included in the case of equivocal results from the parental (P) or first generation (F_1). However, one might question why the study should be extended to a second generation when already clear reproductive and/or developmental toxicity is noted during the first generation. This has been argued to only confirm results already seen and would rarely, if ever, provide significant new information leading to change in regulatory decisions, e.g. setting the overall NOAEL for risk assessment or discerning between category 2 or 1B for classification and labelling, as was published by two research groups [Janer 2007, Martin 2009, Piersma 2011, Rorije 2011] and reported as OECD document no. 176 [OECD no. 176, 2012]. From this can be concluded that a second generation will not change the risk assessment or classification and labelling significantly. However, given the choice of triggering, category 2 reprotoxicants will often trigger the F_2 generation, as doubt may be present whether the substance should be a category 1B cohort. We might even say; will the option of an F_2 generation eventually make the category 2 classification obsolete for chemicals tested in an EOGRTS?

The determination of triggers for a second generation as presented in OECD GD 117 contradicts with the mentioned, convincing literature data on the relevance of the second generation.

Importantly, OECD GD 117 states "When a maximum of 20 EOGRTS have been conducted with an assessment of the second generation, the available results and appropriate scientific evidence will be analyzed. Subsequently, the guidance document [OECD GD 117] will be reviewed and, if necessary, revised in light of experience gained." Overall, one might appreciate that the addition of 20 EOGRTS would add just a small part to the evaluation of the hundreds already available multigeneration studies.

There is also concern that a second generation mating based on first generation triggers has the potential to miss signals of developmental or reproductive toxicity, especially if a test substance produces transgenerational effects that accumulate with time, such as the case with certain epige-

netic changes and endocrine disruptors. The effects on the first generation may not be of sufficient magnitude to trigger an evaluation of a second generation. For this reason, some assert that a second generation mating should be made automatic rather than a triggered event. Certainly, if previous data for a test substance are equivocal or suggest potential transgenerational effects, then it is recommended that a second generation mating be included as part of the EOGRTS design.

The OECD 443 [EU B.56] indicates that it is possible to omit the DNT and DIT cohorts based on existing knowledge of a chemical as well as the needs of the regulatory authorities. However, the information requirements for REACH go a step further as REACH [EU 2015/282] defines that a developmental neurotoxicity or immunotoxicity cohort should only be performed if a particular concern on (developmental) neurotoxicity or (developmental) immunotoxicity is raised from the available data justified by any of the following:

- Existing information on the substance itself derived from relevant available *in vivo* or non-animal approaches, or
- Specific mechanisms/modes of action of the substance with an association to (developmental) neurotoxicity and/or immunotoxicity, or
- Existing information on effects caused by substances structurally analogous to the substance being studied, suggesting effects or mechanisms/modes of action.

Bear in mind that neurotoxicity or immunotoxicity effects observed in adults will also trigger a DNT or DIT cohort, respectively. Information from (validated) non-animal approaches is also mentioned as trigger for these cohorts in the ECHA guidance [ECHA R.7a, 2015]. For example, in the case that an analogue substance does give e.g. cholinesterase inhibition, but *in vitro* testing of cholinesterase activity of the substance itself does not show any activity. The ECHA guidance may be interpreted not to trigger the neurotoxicity cohort if the identified trigger as, in this case, is contradicted by other findings.

The same applies for the immunotoxicity cohort. For example, when an analogue substance shows an affected natural killer-cell function, but an *in vitro* test with the substance itself does not show an effect on the natural killer-cell activity.

Moreover, it is well known that developing offspring can be more sensitive to effects on the nervous or immune systems than adult animals; therefore, the idea of using neurotoxicity or immunotoxicity data obtained from previous adult animal studies to trigger the inclusion of DNT or DIT cohorts in an EOGRTS study is considered by some to be flawed in concept.

Additional experience will be gained in the coming years to ultimately come up with more defined triggers.

Notably, also when the DNT and/or DIT assessments are excluded in the study design, all selected animals, including those in Cohort 2 (DNT) and Cohort 3 (DIT), should be kept until sexual maturation to ensure that sufficient animals (3/sex/litter/dose) are available for evaluation of critical endpoints [OECD GD 151].

6.4. KEY CONSTITUENTS OF THE EOGRTS

This chapter will describe essential details on the four main aspects of the EOGRTS: general toxicity, reproductive toxicity, developmental neurotoxicity, and developmental immunotoxicity.

6.4.1. General toxicity assessment within the EOGRTS

The EOGRTS includes extensive examinations in general toxicity (e.g. clinical pathology and full histopathology) for the adult animals (P and F_1 generation). Labs experienced with general toxicity studies will be able to perform these measurements, however historical control data on these endpoints for lactating rats should be available.

In addition, inclusion of the following simple investigations in the EOGRTS: 1) weekly arena observations, 2) functional observation battery at the end of treatment, and 3) ophthalmoscopic examination at the end of treatment, could eliminate the need for a separate 90-day toxicity study [OECD 408; EU B.26; OPPTS 870.3100], and thereby reduce the total number of animals required. However, aspects as exposure duration, dose levels and route should be considered.

Conversely, it could be argued that if a 90-day toxicity study has already been performed, it could be considered to exclude several P-generation parameters (e.g. clinical biochemistry and macroscopic examination, organ weights and histopathology of all organs) but not the corresponding F_1-generation parameters from the EOGRTS. Necropsy could then be limited to the sex organs and target organs only, provided the same rat strain was used. However, in that case no direct comparison of general systemic effects and reproductive parameters is possible (as was also the case in the past), which might re-introduce the previous issues for determination of secondary effects. Although the possible relationship between general toxicity and reproductive toxicity could be less clear, general toxicity seen in a 90-day toxicity study and reproductive toxicity seen in a Two-Generation Study have been linked before. Moreover, this reduction in F_0 parameters would make the study slightly less complicated and would reduce the costs substantially.

However, as the OECD 443, OECD GD 151 and ECHA Guidance Chapter R.7a.6.4 (reproductive toxicity) [ECHA R.7a, 2015] do not mention this possibility, this would require further discussion and acceptance by the regulatory authorities before being considered. Therefore, the authors of the present manuscript encourage researchers and regulators to explore these proposals. It must be mentioned that this should be carefully considered taking into consideration all the data available and on a case-by-case basis. In the future, when a number of EOGRTS have been conducted, then

the appropriate scientific evidence will be analyzed and the guidance documents may be revised in light of experience gained.

6.4.2. Reproductive toxicity (Cohort 1) assessment within the EOGRTS

The reproductive cohort of the study (Cohort 1) resembles the Two-Generation study, but is divided into two groups of animals: Cohort 1A and 1B. The EOGRTS provides flexibility as the necessity for mating the F_1 animals to produce the second generation is based on regulatory criteria (e.g. for REACH) or triggers obtained during the first generation (e.g. OECD GD 117). OECD 443 mentions "Decisions on whether to assess the second generation should reflect existing knowledge for the chemical being evaluated, as well as the requirements from various regulatory authorities. Procedure for the decision on the internal triggering for producing a 2nd generation is described in OECD Guidance Document no. 117 [OECD GD 117] for those regulatory authorities using internal triggers." It should be noted that OECD GD 117 is only implemented for the US and Canada and not approved by any other OECD member state.

The decision to mate if internal triggers are used has three inherent challenges requiring 1) planning and dedicating resources for rooms and personnel, 2) a rapid and thorough interpretation of the data, and 3) a review and discussion by involved parties to expedite a consensus. As already indicated by Foster and by Fegert and colleagues [Foster 2014, Fegert 2012], all results for making the decision based on internal triggers become available at a very late stage (particularly P reproductive organ histopathology and F_1 estrous cycle data). This leaves only a limited period for review and discussions prior to making a final decision. It should be emphasized that it is of utmost importance that only experienced study directors with a strong scientific support team are appointed because once it is decided not to continue, there is no possibility to later rectify this.
OECD GD 117 is very informative as it mentions all potential effects which may be used as internal triggers; however, the final decision may be subjective and challenging. For instance, the guidance document states that F_1 animals should be mated in the case of biologically relevant decreased F_1 pup body weights in the absence of maternal body weight decrements. What criteria should be applied to demonstrate biological relevance? Should there be statistical significance? What if only a few litters are affected? These considerations should be established on a case-by-case basis and evaluated in great detail.

Recently, the ECHA guidance document on information requirements and chemical safety assessment [ECHA R.7a, 2015] has been updated. The final version of October 2015 as default includes a 10-week premating exposure duration, which can be shortened maximally down to two weeks based on substance specific justifications and e.g. when extension of Cohort 1B is included. The ECHA guidance [ECHA R.7a, 2015] argues that because exposure during the full spermatogenic period (70 days in rats) and folliculogenesis (62 days in rats) are not covered at the time of mating, effects at earlier stages of spermatogenesis and folliculogenesis cannot be reflected in the functional fertility examination.

In contrary, OECD 443 states that "a 2-week pre-mating treatment for both sexes is considered adequate in most cases". According to OECD GD 151 a two-week premating period is sufficient as:

- For most substances steady state concentrations are reached within 2 weeks.
- Spermatozoa acquire motility when they transit the epididymis, which takes two weeks in the rat. Moreover, histopathology and sperm analysis at necropsy will be more sensitive than male fertility and will identify a testicular toxicant.
- A two-week exposure period covers 3 estrous cycles.
- The complete treatment period of the F0-males covers at least 10 weeks.

The pre-mating exposure period was shortened in the OECD 443 because it was recognized that other endpoints (e.g., histopathology, sperm parameters, etc.) are more sensitive than fertility or reproductive performance in rats. However, for industrial chemicals, effects on fertility are one of the criteria used for classification and labeling. Therefore, while reproductive performance is not the most sensitive endpoint, regulators currently require an assessment of fertility after a 10-week pre-mating period in order to satisfy legal requirements to determine whether the compound should be classified as "affecting fertility".

6.4.3. Developmental neurotoxicity (Cohort 2) assessment within the EOGRTS

The developmental neurotoxicity (DNT) phase of the study is subdivided into Cohort 2A with neurobehavioral testing (including functional observation battery, motor activity and acoustic startle habituation) followed by detailed neurohistopathological assessment in adulthood, and cohort 2B with neuropathology assessment at weaning.

The functional observational battery and motor activity tests are included as standard testing in several study types for general toxicity, though acoustic startle response assessment is not. Testing laboratories must ensure these neurobehavioral tests are standardized and are conducted according to pre-defined scales with a high level of inter-observer reliability [Tyl 2008]. Positive control data are also required to demonstrate that testing laboratories can produce reliable data that accurately capture sometimes subtle changes in nervous system function [Crofton 2008]. Importantly, there may be demonstrable behavioral differences between rat strains in response to positive control agents; testing laboratories should be aware of this and have positive control data from the corresponding strains and ages that will be used [Zmarowski 2012]. These are among other strain differences that should be considered when designing this study (see paragraph on culling). Labs experienced with DNT studies [OECD 426; EU B.53; OPPTS 870.6300] can easily incorporate the requested measurements; however, such labs are in the minority.

Neuropathology is required on all control and high dose animals of cohort 2A (between postnatal days (PND) 75-90) and 2B (PND 21-22), and requires additional expertise from testing laboratories. Peripheral nervous system investigation is only required for cohort 2A animals, and plastic embedding of peripheral nerves is strongly encouraged [OECD 426; EU B.53; OPPTS 870.6300].

Morphometric (quantitative) evaluations should be performed on representative areas of the brain and may include linear and/or areal measurements of specific brain regions. More details can be found in OECD 426 [EU B.53; OPPTS 870.6300]. This requires the collection of at least three consecutive sections at each neurological landmark (level) to ensure homologous and representative sections for each specific brain area. Detailed procedures for trimming rat brains are published [Garman 2015]. 'Homologous' in this context implies that the sections are not only bilaterally symmetrical, but are also oriented correctly in the perpendicular axis. The morphometric analysis and necessity of obtaining homologous sections provides the biggest challenge of this cohort and a high level of precision and expertise of the technical staff is required when taking these sections at the appropriate Bregma level [Paxinos and Watson 1998] in a homologous manner, to ensure validity of the morphometric data. Non-homologus (slanted) sections will have altered, typically increased, linear measurements.

The morphometric measurements may be collected from brain sections that are routinely stained with hematoxylin and eosin, but the delimiting structures for the measurements are more defined when the sections are stained with a cresyl violet stain, often performed with a luxol-fast-blue counterstain (i.e. 'Klüver-Barrera stain') [Klüver and Barrera 1953; Duffell 2000]. Cresyl violet is used for staining of the neurons in the brain and spinal cord (to demonstrate the nissl substance in the neurons and cell nuclei) and luxol-fast-blue is used for staining of the myelin.

Linear measurements should be taken from three sections of brain representing forebrain, midbrain and hindbrain. In the forebrain the horizontal anterior commissure provides a useful landmark for ensuring consistency in sectioning, as do the "wine glass" shaped appearances of the lateral ventricles. The midbrain section should be at the level of the optic chiasm, rostral to the pituitary, and the "duck bill" appearance of the dentate gyrus is a useful guide for ensuring consistency between sections. The hindbrain section includes cerebellum and medulla oblongata. This section can be taken as a cross or sagittal section, but if sagittal, the section should be taken just lateral to the midline to keep the medulla oblongata attached to the cerebellum.

Our laboratory includes the analysis of approximately 6 – 8 bilateral linear measurements from the three brain sections; however, the number of measurements and the placements for those measurements may vary [Garman 2015]. The linear measurements should at least include cerebellar height, thickness of the corpus callosum, and thickness of the striatum, cerebral cortex, hippocampus and cerebellar molecular layer [Bolon 2006, Garman 2015]. See Figure 6.2 for the procedures used at WIL Research. Morphometric software, such as Visiopharm software (Visiopharm®, Agern Alle', Denmark), or manual procedures with an ocular reticle may be used in data collection. The measurements should be taken from digital images of scanned histologic slides, using either an automated scanner (e.g., Hamamatsu NanoZoomer) or a conventional flat-bed scanner.

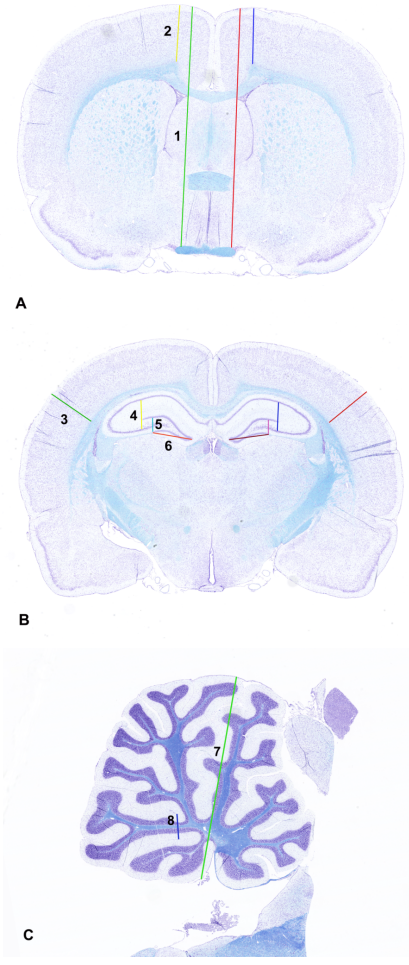


Figure 6.2. Example of 8 quantitative bilateral linear measurements in rat brain at PND 72, and stained with cresyl-violet/ luxol fast blue.

A) The forebrain section of cerebral cortex includes measurements 1 and 2. Measurement 1 is the greatest vertical height of the hemisphere tangent to the most medial part of the lateral ventricle. Measurement 2 is the vertical cortical thickness of the cerebral cortex taken at the apex of the corpus callosum and parallel to the midline. Notice the anterior commissure and the "wine glass" appearance of the lateral ventricles, which both provide useful landmarks.

B) The midbrain section includes measurements 3 - 6. Measurement 3 is the radial thickness of the fronto-parietal cortex at the 2 o'clock and 10 o'clock positions. Measurement 4 is the vertical height between the layers of the hippocampal pyramidal neurons along a line that passes through the termination of the dorsal limb of the medial dentate hilus. Measurement 5 is the vertical height of the dentate gyrus measured at the termination of the ventral limb. Measurement 6 is the length of the ventral limb of the dentate gyrus. Note the resemblance of the contour of the dentate gyrus to that of a "duck bill".

C) The hindbrain (sagittal section of cerebellum) includes measurements 7 and 8. Measurement 7 is the height of the cerebellum taken from the heel of lobule 10 to the superior portion of lobule 6. Measurement 8 is the distance across the base of cerebellar lobular 9 measured perpendicular to the white matter tract.

There is evidence that gross dimensions and weight of the rat brain undergo changes during immersion fixation with formalin. While STP "best practice" recommends perfusion fixation at both PND 22 and PND 70 [Bolon 2006], immersion fixation is permitted [OECD 426; OECD 433]. Our internal studies indicate rat brains that are fixed by immersion in neutral-buffered formalin exhibit an immediate increase in weight, width and length as compared to the weight and dimensions of the unfixed brain [Acciani 2015]. All the measurements (weight, width, and length) may be considered stable at and after the sixth day post-immersion (Table 6.2). Therefore, in DNT studies if immersion fixation is used, it is important to i) perform histologic sampling and processing at equivalent times for all rats; ii) randomize the date of histologic sampling and processing across all treatment groups; and/or iii) section brains on or after day 6 of fixation.

Table 6.2. Day of stabilization of weight, width and length of rat brain after immersion-fixation in formalin.

Measurement	Males	Females
Weight	Day 6	Day 6
Width	Day 4	Day 1
Length	Day 3	Day 2

It should be noted that OECD 443 mentions that perfusion fixation of Cohort 2B animals is optional. However, as morphometric evaluations should be performed on representative areas of the brain, we strongly advise to incorporate this procedure for both ages.

It is worthy to note that the position of the guidance document is somewhat contradictory on whether the DNT assessment within the EOGRTS provides sufficient assessment of nervous system development as a stand-alone DNT study. The guidance document [OECD GD 151] mentions in paragraph 100 "The neurotoxicity testing in OECD 443 aims to provide an initial assessment of neurotoxicity potential but does not include all facets of a complete DNT study. Thus it is not intended as a replacement for a DNT study, nor is it appropriate to interpret the results from the OECD 443 DNT assessment as a replacement for conducting a DNT study where one may be required." The adequacy of the DNT portion of the EOGRTS will likely depend on regulatory decision-making taking into account the mechanism of action and results of prior studies. In contrast, the same guidance document mentions in paragraph 13 "The number of animals used will be reduced especially when the EOGRTS is considered to fulfill information requirements on reproduction toxicity, DNT and DIT, as compared to the number of animals used in the three separate studies for these endpoints".

Comparing these guidelines, there are no requirements for learning and memory testing in the OECD 443 and more extensive examinations are required for the peripheral nervous system in the DNT study. Under REACH the DNT cohort may be replaced by a separate DNT study if considered more applicable [EU 2015/282]. As the EOGRTS requires extensive evaluation of the nervous system, learning and memory data do not impact the LOAEL as reported by Raffaele et al. [Raffaele

2010] and the EOGRTS includes most parameters required by the DNT study, in general we take the approach of not performing an OECD 426 [EU B.53; OPPTS 870.6300] unless there are equivocal findings in the DNT cohort that require further evaluation in a full DNT study.

6.4.4. Developmental immunotoxicity (Cohort 3) assessment within the EOGRTS

The developmental immunotoxicity (DIT) portion of the study (Cohort 3) consists of a T-cell dependent antibody response assay (TDAR), which can easily be incorporated in labs with immunotoxicity experience.

Toxicology studies often refer to historical control data (HCD) in addition to the concurrent control data. The maturation of the immune system in rodents is considered to be achieved from 4 weeks of age in the rodent [Landreth 2002]. So HCD combining 6-10 weeks of age and covering *per se* 8 weeks of age (i.e. age of examination in EOGRTS) corresponds to a mature immune system allowing proper antigen recognition to occur. When immunological parameters and functions are considered, large individual variability and sensitivity are often encountered, in addition to inherent variations in used techniques, reagents and equipment (see Figures 6.3 and 6.4 for historical data). This high variation can result in a large range of extreme values within the historical data which may lead to misinterpretation of results because values typically "remain within the historical control data range". It is therefore of utmost importance to have an antigen-naïve pretest value for each individual to properly assess the TDAR response. This approach takes into account any potential somatic recombination which could lead to pre-existing antigen-cross-reactivity antibodies. In addition this also allows validating the sensitivity and the specificity of the ELISA kit used for IgM measurement. Since the negative control animals will all be positive with anti-KLH IgM (when no TDAR inhibition occurs), the antigen-naïve samples allow demonstrating that the ELISA method worked properly with negative (antigen-naïve) values in the study animals, re-enforcing the robustness of the assessment. Remarkably, this crucial aspect is not mentioned in the guideline or guidance document.

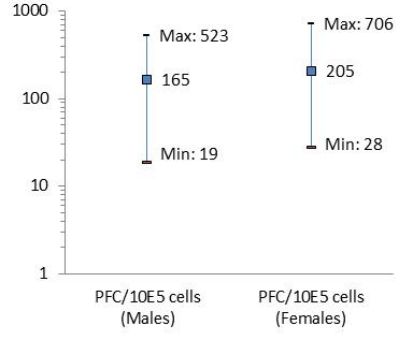


Figure 6.3. Historical data for anti-SRBC PFC units in rat spleen (N = 196, 6-10 weeks of age). Mean with the maximum and the minimum individual values.

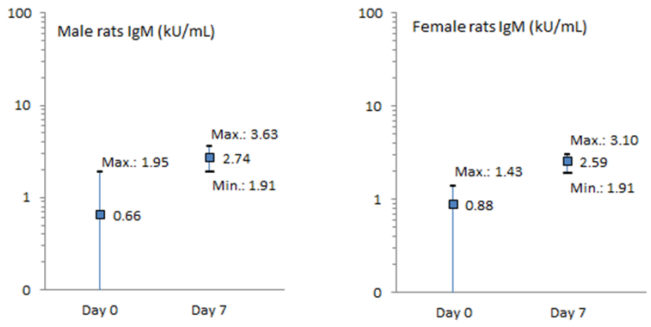


Figure 6.4. Historical data for anti-KLH IgM in rat (N = 30, 6-10 weeks of age). Before the antigenic challenge (Day 0) and seven days after the challenge (Day 7). Mean with maximum and minimum individual values.

OECD 443 mentions that both sheep red blood cells (SRBC) and keyhole limpet haemocyanin (KLH) are suitable antigens. SRBC have been historically used and are still requested in current US EPA testing guidelines for assessment of immunotoxicity in adult animals [OPPTS 870.7800]. However, as experience with KLH increases, this immunogen has a longer shelf life (useful in this study since the animals included can be spread over a 2-week period), KLH responses are less variable between batches as it is commercially available as GMP (Good Manufacturing Practices) material, and KLH is more readily available (as opposed to SRBC, an immunogen that has variable immunogenic potency and is not available as a standardized reagent), we consider KLH better suited for the EOGRTS. In addition, when needed for a more thorough evaluation, both primary (IgM) and secondary (IgG) responses can be studied (instead of IgM only as is requested in the guideline) when using KLH.

The impairment of the TDAR response may be evaluated by counting specific plaque-forming cells (PFC) in the spleen or by determining the titer of SRBC- or KLH-specific IgM antibody in the serum by ELISA (Enzyme-Linked ImmunoSorbent Assay).

There are many advantages of measuring serum antibodies by ELISA when compared to the PFC assay. The ELISA technique is readily standardized, uses small volumes of blood, microplate-kits are commercially available, allows automatic measurement using micro-plate readers, allows assessment of in vivo total antibody formation versus spleen-specific responses, and the serum samples can be stored before analysis (providing flexibility) and can be retested for confirmatory purpose. Measuring antibodies by ELISA is technically easier and less labor intensive than the PFC assay [Herzyk and Bussiere 2008]. In addition, each animal is its own negative control (pre-challenge value) and therefore does not require additional animals in the negative (vehicle) control compared with the splenic IgM PFC assay [Gore et al. 2004].

Therefore, ELISA is considered superior to the PFC assay and KLH a more reproducible antigen to use in TDAR than the SRBC.

Table 6.3.a. Endpoints and decision points in the EOGRTS (when Cohorts 2 and 3 are included and a 2-week premating period is used).

Generation	Phase	Study week	Timing	N	Who	No. of animals over time
P	Start treatment	0	-	200	All	200 (25/sex/group)
	Estrous cycle determination	0	For 2 weeks	100	Females	
	Start mating	2	-	200	All	
	Delivery of litters	5	-	100	Litters	
	Anogenital distance	5	PND 1	100	Litters	
	Culling of pups (possible T4, TSH)	6	PND 4	100	Litters	
	Areolae/nipple retention male pups	7	PND 13	100	Litters	
	Assignment of F_1-generation[a]	8	PND 21	-	-	
	T4, TSH	8	PND 21	80	Surplus pups	
	Limited necropsy	8	PND 21	80	Surplus pups	
	Clinical pathology (H+B+U[b], T4, TSH)	8	Lactation day 22	40	Females	
	Full necropsy	8	Lactation day 22	100	Females	
	Full histopathology P-females	-	-	50	Group 1+4	
F_1	Limited necropsy incl. perfusion	8	PND 21-22	80	Cohort 2B	480 (60/sex/group) + 100 (25 P-males/ group)
	Histopathology brain incl. morph.	-	-	40	Cohort 2B	
	Auditory startle	8	PND 24 ± 1	80	Cohort 2A	
	Check for vaginal opening (VO)	8	From PND 25 onwards	240	All females	
	Determination day of first estrus	9	From positive VO onwards	80	Cohort 1A	
	Check for balanopreputial separation	10	From PND 35 onwards	240	All males	
(P)	Clinical pathology (H+B+U[b], T4, TSH)	10	After 10 weeks of treatment	40	P-Males	
(P)	Full necropsy (incl. sperm exams) P-Males	10	After 10 weeks of treatment	100	P-Males	480 (60/sex/group)

(P)	Full histopathology P-males	-	-	50	P-Group 1+4	
	Blood sampling from antigen-naïve animals	12	PND 51 ± 3	80	Cohort 3	
	KLH immunization	12	PND 51 ± 3	80	Cohort 3	
	TDAR assay and limited necropsy	13	PND 56 ± 3	80	Cohort 3	
	Extensive functional observation battery	14	PND 63-75	80	Cohort 2A	400 (50/sex/group)
	Estrous cycle determination	16	From PND 75 for 2 weeks	80	Cohort 1A	
	Limited necropsy incl. perfusion	16	PND 76-90	80	Cohort 2A	
	Histopathology CNS+PNS incl. morph.	-	-	40	Cohort 2A	320 (40/sex/group)
	Clinical pathology (H+B+U[b], T4, TSH)	18	≥ PND 90	80	Cohort 1A	
	Full necropsy (incl. sperm exams)	18	≥ PND 90	160	Cohort 1A	
	Splenic lymphocyte subpopulation analysis	18	≥ PND 90	80	Cohort 1A	160 (20/sex/group)
	Full histopathology incl. follicle counts and sperm staging	-	-	80	Cohort 1A Group 1+4	
	Limited necropsy[c]	19	≥ PND 97	160	Cohort 1B	0

[a] On PND 21, pups from all available litters (4 pups/sex/litter) up to 20 per group will be selected and assigned according to the following schedule:
Cohort 1A (Reproductive): 1 male and 1 female pup
Cohort 1B (Reproductive): 1 male and 1 female pup
Cohort 2A (Neurotoxicity): 1 male or 1 female pup
Cohort 2B (Neurotoxicity): 1 male or 1 female pup
Cohort 3 (Immunotoxicity): 1 male or 1 female pup
Surplus (Spare animals): 1 male or 1 female pup
[b] H+B+U = hematology, biochemistry and urinalysis
[c] If not triggered for a second generation of mating.

Table 6.3.b. Endpoints required for the triggered second generation.
When mating for a second generation will be needed, necropsy in Week 19 of study will not be performed, and then the study will be extended with the following:

Generation	Phase	Study week	Timing	N	Who	No. of animals over time
F_1 Cohort 1B	Start mating	19	PND 90-120	160	All	160 (20/sex/group)
	Estrous cycle determination	19	During mating	80	Females	
	Delivery of litters	22	-	80	All	
	Anogenital distance	22	PND 1	80	All	
	Culling of pups	23	PND 4	80	All	
	Limited necropsy	23	After delivery litters	80	Males	
	Areolae/nipple retention	24	PND 13	80	Male pups	80 (20 females/group)
	Limited necropsy	25	PND 21	80	Females and pups	0

Nevertheless, since the quality and reliability of the numerous kits available on the market are very heterogeneous, it is strongly advised to evaluate the intended kit in a challenging validation process before selecting a routine tool. This should be done for every change in manufacturer of the kit or in the kit itself (e.g., polyclonal antibodies vs. monoclonal antibodies).

OECD 443 mentions that performance of the TDAR method should be confirmed as part of the optimization process by laboratories setting up the assay for the first time, and periodically (e.g. yearly). And that additional pups may be required from the control group to act as positive control animals for the TDAR. In case more EOGRTS are performed during the same period, instead of using control pups from the ongoing experiment, we recommend running a separate positive control assay as it renders the ongoing EOGRTS slightly less complicated. However, when EOGRTS are performed less frequently, from an ethical perspective it would be better to use control pups from the running test.

6.5. LOGISTICS

The logistics of an EOGRTS, which includes both cohorts and the possibility to extend with a second generation, is not like any other guideline study. The study types that are most comparable are the developmental neurotoxicity studies [OECD 426; EU B.53; OPPTS 870.6300] for industrial chemicals and agrochemicals and juvenile toxicity studies for pharmaceuticals. However, both of these can be scheduled over a broader range as these studies do not generally include a premating period. Due to the premating period in the EOGRTS, mating of all animals should start on the same day. According to WIL Research's experience, this means that at least 95% of the female rats (~N=95) will mate in the first 4 days, and therefore all determinations that are scheduled on a fixed age, need to be performed in those 4 days. Especially, the period of PND 21-22 is extremely demanding with numerous activities for many animals (i.e. assignment into cohorts, thyroid hormone analyses and necropsy of surplus animals, clinical pathology and full necropsy of P-females, necropsy and perfusion of Cohort 2B pups).

OECD GD 151 allows for the possibility to conduct the EOGRTS in blocks (e.g., parental animals divided into two or three groups with a staggered study start – e.g. 1 week between blocks) for some laboratories that may not have the capacity to perform all examinations at the same time. This approach may allow for easier animal management and data collection. Performing the EOGRTS in blocks reduces the number of animals evaluated on the same day and may potentially decrease diurnal variation at the time of testing, which may be greater for animals with mating starting on the same day than for smaller numbers of animals evaluated in blocks on different days. Consideration should be given, based on a laboratory's resources, whether to perform the EOGRTS in blocks. Care should be taken to minimize potential variation introduced by temporal differences in data collection and to ensure that the data from each block can be efficiently obtained and evaluated so as not to delay the decision of triggered events (when internal triggers are used). If the EOGRTS is conducted in blocks, there may be a greater obligation on the part of the researchers to examine blocks for any temporal effects using statistical approaches.

Table 6.3.a provides the detailed planning of an EOGRTS including Cohorts 2 and 3. Table 6.3.b describes the additional activities when mating for a second generation is required.

6.6. DISCUSSION

This chapter will describe five additional considerations that relate to the EOGRTS: culling of the litters, extent of the preliminary study for dose level selection, adjustment of dietary exposure levels, type of diet and bedding, and an example of an alternate design.

6.6.1. Culling (reduction of litter size)

The choice of rat strain should be carefully considered in light of all available information. OECD 443 states "On day 4 after birth, the size of each litter may be adjusted by eliminating extra pups by random selection to yield, as nearly as possible, five males and five females per litter." This would result in 30 reserve pups/sex/group, which might be considered excessive. Moreover, culling to 4 pups/sex/litter would still give 10 reserve pups/sex/group. In some strains that produce smaller or more variable litter sizes (e.g. Wistar Han), it can be problematic to obtain a sufficient number of litters with 5 pups/sex/litter. This should be taken into account when designing an EOGRTS.

OECD GD 117 also addresses this issue and states that the target adjusted litter size should be based on the normal litter size of the strain used. Ten pups per sex is generally appropriate for Sprague Dawley rats with a natural litter size of about 14, but 8 per litter is more achievable for Wistar rats which normally have smaller litter sizes. Historical control data from WIL Research is provided in Table 6.4.

Table 6.4. Overview of the size and sex distribution of control litters for Wistar Han and Sprague Dawley rats obtained at our laboratories from representative studies performed in 2008 to 2014.

	Den Bosch, the Netherlands	Ashland, USA	Lyon, France	
	Wistar Han	Sprague Dawley	Wistar Han	Sprague Dawley
No. of control litters	1778	571	333	397
Mean no. of pups/litter	11.6	14.3	10.7	12.8
Selection of 10 pups/litter				
% of litters 5+5	54%	70%	41%	58%
% of litters 4+6 or 6+4	68%	85%	56%	69%
Selection of 8 pups per litter				
% of litters 4+4	76%	87%	67%	76%
% of litters 3+5 or 5+3	86%	93%	80%	87%

Some researchers debate the scientific basis of culling litters to a pre-determined number of pups per sex, asserting that such a procedure of artificially eliminating pups from a litter may limit the ability to fully assess certain indicators of toxicity that may impact litter size, postnatal survival, or

pup body weight [Palmer and Ulbrich, 1997]. The standardization of litter sizes on PND 4 to potentially reduce variance and decrease confounding factors related to variability in litter size remains a common and accepted practice in reproductive toxicity studies including the EOGRTS, but there are some who question the utility of this procedure and reasonable consideration should be given when designing these studies to help ensure consistency in procedures across studies, minimize potential sources of variability, and make sure that an adequate number of pups are retained to meet the scientific objectives of the study and maximize the amount of information obtained.

6.6.2. Extent of the dose range-finding study

Guidance Document no. 151 strongly recommends performing a dose range-finding study, including endpoints of mating, fertility, litter production and pup survival. The use of dose range-finding studies is recommended to help ensure that the dose levels do not impact fertility in the absence of significant systemic toxicity, such that insufficient pups are produced for assessment of the F_1 generation, and to aid in dose level selection based signs of systemic toxicity in conjunction with a toxicokinetic assessment of exposure, if warranted.

EU Regulation 2015/282 [EU 2015/282] states that fertility is the focus of the EOGRTS and dose selection should be appropriate to meet risk assessment and classification and labelling purposes for fertility. According to the ECHA guidance R.7a [ECHA R.7a, 2015] selection of the highest dose level based merely on toxicokinetic data as mentioned in OECD 443 and OECD GD 151 may not allow comparison of adverse effects on fertility with systemic toxicity, i.e. discern between a direct fertility effect and a secondary fertility effect, as part of the evaluation for classification and labelling. Based on an evaluation of our Dutch laboratory's database of 134 reproduction/developmental screening (i.e. OECD 421 and 422 studies [OECD 421 and 422]) [Beekhuijzen et al. 2014], dose levels for an OECD 421 repro screening study were based on a previous 28-day toxicity study and dose levels for an OECD 422 combined 28-d/repro screening study were selected based on the results of a dose range-finding study (ranging from 10 to 14 days of treatment) in which mating was not included. Out of the 134 screening studies evaluated, only one OECD 422 study showed developmental and reproductive toxicity in all dose groups with a clear dose-related trend. For all the other screening studies the dose levels were adequately set to produce some toxicity at the highest dose level based on repeated dose toxicity data without mating. Taken together, at the time an EOGRTS needs to be performed, one must not overlook the possibility that a sufficient amount of data may be available from previous guideline studies (e.g. 90-day toxicity study, repro screening study) to adequately select dose levels. Moreover, with the recently updated OECD 421 and 422 guidelines (e.g. additional endocrine disruptor endpoints and extended postnatal follow-up), even more relevant information would be available.

Another reason that GD 151 [OECD GD 151] recommends performance of a dose range-finding study (including mating and litter production) is to obtain information on pup exposure by assessing test substance concentration in pup blood and maternal milk samples. Pups can start to eat diet beginning late in the second/early third week of postnatal development. If pups are not

exposed to the test substance via milk or direct dosing prior to the time that they are able to consume dietary admixes containing the test substance (in dietary studies) or until direct dosing commences (in gavage studies) typically at, or shortly after weaning, there is a gap in exposure during an extremely critical window of development. It may therefore be necessary to consider direct dosing of pups during the pre-weaning period.

Given the complexity and duration of an EOGRTS, we would strongly recommend that a dose range-finding study will be performed (with mating and litter production if there are no data from previous studies), including toxicokinetic (TK) assessments, especially for assessing pup exposure, when warranted.

The potential value of including TK assessments for selecting dose levels and interpreting data is emphasized by an EOGRTS study with 2,4-Dichlorophenoxyacetic acid (2,4-D) conducted by Marty et al. [Marty 2013]. The effects on thyroid endpoints in the high-dose females during gestation were associated with non-linear toxicokinetics, resulting in high plasma 2,4-D levels due to saturation of renal clearance. These effects were not seen at 2,4-D doses in the linear toxicokinetic range.

The authors agree with the following statement of the OECD 443: “although not required, TK data from previously conducted dose range-finding or other studies are extremely useful in the planning of the study design, selection of dose levels and interpretation of results. Of particular utility are data which: 1) verify exposure of developing fetuses and pups to the test compound (or relevant metabolites), 2) provide an estimate of internal dosimetry, and 3) evaluate for potential dose-dependent saturation of kinetic processes.”

6.6.3. Adjustment of dietary exposure levels

The preferred route of exposure is dietary. The amount of test substance in terms of ppm is generally kept at a constant level during the study; however, during lactation there is a significant increase in exposure based on mg test compound per kg body weight as food consumption during this period more than doubles to facilitate milk production. To limit overexposure of the offspring during such critical windows of development and in accordance with guidance by Cooper and colleagues [Cooper 2006], dietary concentrations can be adjusted to account for the large and rapid increase in feed consumption typical for rats in mid- to late lactation [Carney et al. 2004]. The authors would recommend these dietary adjustments to maintain more consistent doses in terms of mg/kg body weight across study phases; this is however not mentioned in OECD GD 151. Table 6.5 gives the adjustments in ppm during the lactation period based on historical control data of relative food consumption for Wistar Han and Sprague Dawley rats.

Table 6.5. Adjustments in dietary concentrations throughout the different phases of a generation study applied at WIL Research for Wistar Han rats and Sprague Dawley rats.

	Concentration in diet (ppm)	
Study period	**Wistar Han**	**Sprague Dawley**
Days 1-4 of lactation	67%	57%
Days 4-7 of lactation	50%	44%
Days 7-21 of lactation	40%	32%
Remainder of study	100%	100%

It is also worthy of note that there will be a contribution from the pups with respect to food consumption and test article intake, particularly during the last week of lactation (Days 14-21) because the pups also consume diet during this period [Hanley and Watanabe 1985; Yoon and Barton 2008]. Moreover, pups do not achieve adult levels of feed intake/kg body weight until well after weaning (~PND 70) [Marty et al. 2013]. Therefore, even with dietary adjustments during lactation, F_1 offspring will receive greater dose levels of test compound than P adults. This is important when evaluating results in the F_1 generation.

6.6.4. Type of diet and bedding

OECD 443 recommends the use of standardized open-formula laboratory animal diets in which oestrogenic substances have been reduced. An analysis carried out during the OECD validation of the uterotrophic assay showed that whilst high levels of phytoestrogens (> 325-350 µg/g total genistein equivalents; TGE) could diminish the responsiveness of this very distinctive assay, levels below this had little effect [Owens et al. 2003].

However, we would not advise to change the type of diet only for the EOGRTS, as historical control data (based on the standard laboratory diet) are crucial for a correct interpretation of the obtained results. In addition, there are no reports of similar effects in apical tests such as the EOGRTS. Furthermore, alteration of dietary ingredients is not encouraged as the loss of key constituents and trace elements might adversely affect parturition and survival of the neonates.

It is therefore advised to request analytical certificates from the supplier for each batch of diet intended for use in an EOGRTS. It will not be needed to use low phytoestrogen diets, but rather to make sure each batch of standard diet contains < 325-350 µg/g TGE.

The same applies for bedding.

6.6.5. Example of an alternate design

The National Toxicology Program (NTP) has proposed an alternate study design [Foster 2014], termed the modified one generation (MOG) reproduction study (see NTP website for detailed study protocol). The MOG study encompasses developmental and reproductive toxicity endpoints as well as the selection of appropriate dose levels for a cancer bioassay through evaluation of target

organ toxicity that is based on test article exposure starting during gestation. In Fosters' paper, the MOG design is compared against the EOGRTS. The MOG study design does not employ the use of internal or external triggers, although it adheres to the principles of the 3R's as it reduces overall animal usage compared to the Two-Generation and 90-day toxicity studies. Mating to produce a second generation is included by default in the MOG design, and used to obtain prenatal development toxicity and detailed fetal morphology data. This design would be of added value for test substances also requiring prenatal development data. However, in the cases where these data are already available, the MOG design would use more animals than the EOGRTS. In addition, currently the MOG design is not included in regulatory information requirements.

6.7. CONCLUSIONS

In this review, we evaluated the hallmarks of the EOGRTS and addressed specific considerations that should be taken into account when considering to implement this study design. The complexity of the study should not be underestimated and experienced testing laboratories with sufficient resources and historical control data for all parameters should conduct the studies. For the coming years, it will be important to gather and review data from the EOGRTS to evaluate the utility of this design for identifying developmental and reproductive toxicants and to further optimize the design if needed. Altogether, this should then also provide more insights on any benefit of an (internal) triggered second generation mating.

Four studies that employed the EOGRTS concluded that, though challenging, the EOGRTS is manageable for facilities experienced in performing Two-Generation reproduction, developmental neurotoxicity and immunotoxicity studies. Logistical bottlenecks occur at certain study phases (e.g. PND 21), but can be resolved with detailed planning and execution by experienced personnel [Fegert et al. 2012].

The EOGRTS improves upon several aspects of the Two-Generation study by including evaluation of more endpoints. The potential use of triggering of the various cohorts further enhances the flexibility and relevance of the study to assess endpoints of concern while minimizing the use of animals to the fullest extent possible. However, the acceptability and utility of triggering, and the ability for the cohorts to provide sufficient information as stand-alone studies still remain the subject of debate. Taken together, the EOGRTS is a definite step in the right direction of an enhanced study design but with the considerable technical, scientific and regulatory challenges it presents, the ultimate design of the ideal study remains a work in progress. We would therefore encourage other laboratories to publish their valuable experiences with implementation and performance of the EOGRTS.

Acknowledgements

The authors thank their colleagues in Lyon (France), Ashland (OH, USA), Hillsborough (NC, USA) and Den Bosch (The Netherlands) for their valuable contribution in creating one optimal EOGRTS design implemented at all global WIL Research sites.

In addition, the authors would like to thank all 8 reviewers, which were selected by the Editor and anonymous to the authors, for their valuable input and helpful remarks. Their constructive comments and suggestions greatly improved the manuscript.

Ethical standards

The animal studies were performed with approval of the local regulatory agencies, and all study protocols complied with the federal guidelines.

Declaration of interest

The authors are employees of WIL Research, a contract research organization. WIL Research routinely conducts studies of the type described in the paper for a wide range of clients and reports are provided that are routinely submitted to regulatory authorities. The authors have not been involved in any regulatory or legal proceedings related to the contents of the paper. The conclusions drawn and recommendations offered are exclusively the professional opinions of the authors.

Chapter 7

Guidance on the Selection of Cohorts for the Extended One-Generation Reproduction Toxicity Study (OECD Test Guideline 443)

Nigel P. Moore[1], Manon Beekhuijzen[2], Peter J. Boogaard[3], Jennifer E. Foreman[4], Colin M. North[4], Christine Palermo[5], Steffen Schneider[6], Volker Strauss[6], Bennard van Ravenzwaay[6], Alan Poole[7]

[1] Ubrs GmbH, Basel, Switzerland

[2] WIL Research Europe BV, s'-Hertogenbosch, The Netherlands

[3] Shell International BV, The Hague, The Netherlands

[4] ExxonMobil Biomedical Sciences Inc., Annandale, NJ, USA

[5] ExxonMobil Petroleum and Chemical, Brussels, Belgium

[6] BASF SE, Ludwigshafen, Germany

[7] European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), Brussels, Belgium

Published in: Regulatory Toxicology and Pharmacology. 2016 Oct;80:32-40. doi: 10.1016/j.yrtph.2016.05.036. Epub 2016 May 28.

ABSTRACT

The extended one-generation reproduction toxicity study (EOGRTS; OECD test guideline 433) is a new and technically complex design to evaluate the putative effects of chemicals on fertility and development, including effects upon the developing nervous and immune systems. In addition to offering a more comprehensive assessment of developmental toxicity, the EOGRTS offers important improvements in animal welfare through reduction and refinement in a modular study design. The challenge to the practitioner is to know how the modular aspects of the study should be triggered on the basis of prior knowledge of a particular chemical, or on earlier findings in the EOGRTS itself, requirements of specific regulatory frameworks notwithstanding. The purpose of this document is to offer guidance on science-based triggers for these extended evaluations.

Keywords
Cohorts, EOGRTS, Extended one-generation reproductive toxicity study, Developmental toxicity, Fertility, Rat, Reproduction, Triggers

7.1. INTRODUCTION

The reproductive cycle in eutherian mammals is a complex, highly-integrated process that may be adversely affected by toxicant acting at different stages of the cycle and through diverse modes of action. While various study designs have been developed to assess the potential effects of chemicals and agrochemicals upon selected parts of the reproductive cycle [ECETOC, 2002], the two-generation reproductive toxicity test - OECD Test Guideline 416 [OECD 416] - has been considered the 'gold standard' for the holistic assessment of effects on reproduction as exposure to a test substance is maintained throughout the cycle.

The two-generation reproductive toxicity test is usually carried out in the rat as the preferred species. Groups of animals are exposed to the test substance for ten weeks prior to mating, throughout the mating period, gestation, and lactation through to weaning, at which point selected individuals of the F_1 generation are maintained on the same exposure protocol to produce the F_2 generation (Figure 7.1.A). The study design is animal-intensive, with approximately 2600 animals being used throughout the course of a standard study [Cooper 2006], and allows evaluation of effects on mating behaviour; fertility; *in utero* and postnatal development, including landmarks of sexual development and maturation; and reproductive tract organs. The test does not, however, evaluate effects on all facets of the reproductive cycle, such as the developing nervous and immune systems. Furthermore, recent analyses have questioned the practical and scientific value of mating the F_1 generation to produce the F_2 generation at a substantial cost in terms of animal use [Janer 2007; Piersma 2011; Rorije 2011].

An alternative study design—the extended one-generation reproduction toxicity study (EOGRTS)—was originally proposed by the ILSI Health and Environmental Sciences Institute's Agricultural Chemical Safety Assessment (ACSA) Technical Committee, as part of a tiered testing approach for agricultural chemicals [Cooper 2006]. This test design incorporated the preliminary assessment of developmental neurotoxicity (DNT) and developmental immunotoxicity (DIT) and removed the mandatory requirement for a second generation breeding. At that time the logic for including DNT and DIT assessments in a one generation study was that for agricultural chemicals DNT was a quasi-mandatory testing requirement, and requests to investigate immune toxicity (including DIT) started to increase. The inclusion of DNT and DIT assessments into the EOGRTS was considered to be a more efficient way to address these testing requirements.

An essential difference between the EOGRTS and the two-generation reproduction toxicity study is the omission of the second breeding. In support of the proposal to remove the second breeding as a standard requirement ACSA cited an unpublished US EPA review of data collected on the F_2 generation from studies of 350 agrochemicals which indicated that, in all except two possible cases, adverse effects would have been identified in the first generation or from other toxicity studies [Cooper 2006]. Additional evaluations of the added value of the F_2 generation breeding have since examined its impact upon hazard identification (classification) and risk assessment (identification

of NOAEL and LOAEL values), as well as the sensitivity of internal triggers from the F_1 generation [Dang 2009; Janer 2007; Myers 2008; Piersma 2011; Rorije 2011]. An analysis of two-generation studies for fifty substances that are classified for effects on reproduction revealed that in only one case was the classification determined by unique effects in the second breeding phase, but that in this case further testing would have been triggered by other available data [Rorije 2011]. Furthermore, in a key analysis of 498 two-generation studies representing 438 substances the second breeding did not drive the derivation of the NOAEL [Piersma 2011].

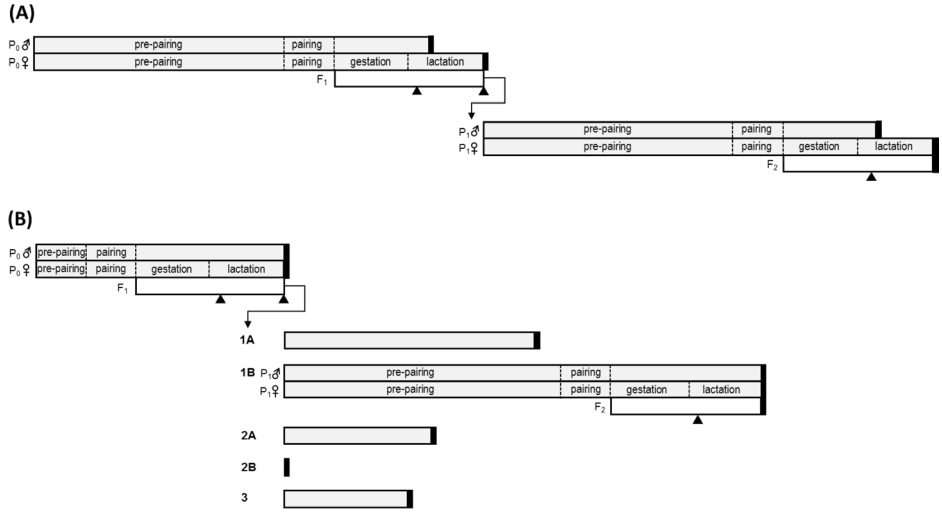


Figure 7.1. Schematic diagrams of (A) the two-generation reproductive toxicity study, OECD TG 416, and (B) the extended one-generation reproductive toxicity study, OECD TG 443.

A) Animals are 5–9 weeks of age at the start of treatment. Treatment of males and females continues for ten weeks prior to pairing; during the pairing period, up to two weeks' duration; and females are then continuously treated during gestation and lactation to three weeks postpartum when the offspring are weaned and the dams are sacrificed. Parental males are sacrificed when no longer needed for assessment of reproduction, usually shortly after parturition when postnatal survival of the offspring is assured. Litter sizes may be standardised at PND 4, and at weaning are reduced to one male and one female, which go on to produce the second generation.

B) Animals are usually treated for two weeks prior to pairing, although this may be extended if additional information or regulatory requirements necessitate; throughout the pairing period; females are then continuously treated during gestation and lactation to three weeks postpartum, when the offspring are weaned and the dams are sacrificed; parental males are continuously treated for a minimum of ten weeks. Litter sizes may be standardised at PND 4, and at weaning the offspring are allocated randomly to the cohorts; one male and one female from each litter are assigned to cohorts 1A and 1B, one male or one female is assigned to each of cohorts 2A, 2B, and 3. If the second generation (F_2) mating is not required, cohort 1B animals are necropsied at approximately 14 weeks of age. If the second generation mating is required then, based on weight of evidence, it may be sufficient to terminate all cohort 1B progeny (F_2 litters) on PND 4. Shaded bars, direct exposure to the test substance; open bars, indirect exposure, although for substances administered in the diet or drinking water offspring do become directly exposed from around the end of the second week postpartum; black bars, necropsy; black triangles, interim litter standardisation at PND 4 (optional) and weaning. P_0, first generation parental animals; F_1, first generation offspring; P_1, second generation parental animals; F_2, second generation offspring.

Based on data suggesting that the second breeding adds little or no critical information to hazard and risk assessment, ACSA proposed that examination of second generation was unnecessary unless triggered by adverse effects in the parental animals or F_1 progeny [Cooper 2006]. Eventually a fully modular study design was proposed in which the DNT and DIT as well as other evaluations would also be triggered if data from studies in parent animals, the F_1 generation, or other toxicological investigations suggested a potential for adverse effects on neurodevelopment or the immune system. If there were no alerts from appropriately designed and robust studies, then such evaluations could be waived rather than being a mandatory requirement [ECETOC, 2008a, 2008b; Moore 2009; Vogel 2010]. The OECD adapted the EOGRTS from that originally proposed by ACSA, to provide flexibility and emphasise the importance of existing knowledge in study design [OECD 443]. The OECD test guideline (TG) describes three cohorts of F_1 generation animals that may be used to address specific developmental endpoints (Figure 7.1B). The conduct of these cohort evaluations are not mandatory to the validity of the test, nor are they default requirements in all regulatory frameworks, such as REACH [EU 1907/2006 corrigendum, EU 2015/282]. Thus OECD TG 443 offers flexibility in the conduct of these evaluations:

> "*Decisions on whether to assess the second generation and to omit the developmental neurotoxicity cohort and/or developmental immunotoxicity cohort should reflect existing knowledge for the chemical being evaluated, as well as the needs of various regulatory authorities*." [OECD 443]

In order to reduce the number of animals and procedures involved in meeting the requirements of specific regulatory frameworks, a set of triggers and waivers for the inclusion or exclusion of evaluations was proposed that allow the EOGRTS to not only *reduce* animal use, but to *refine* their use in terms of limiting unnecessary manipulations [ECETOC, 2008a, 2008b; Moore 2009]. Indeed, the core study design provides sufficient evidence to inform hazard identification for reproductive, DIT, or DNT effects in the progeny where prior knowledge for effects is unavailable. The purpose of this communication is to give further practical guidance on the identification and exercise of triggers and waivers, in order to optimise the reduction and refinement possibilities that the OECD TG 443 presents but without compromising hazard identification and risk assessment. It is important to note that there are different technical standards for developing a waiver compared to a trigger. A trigger is, generally, a clear indication that further testing is needed, whereas a waiver is a compilation of evidence that no further testing is necessary, therefore the data that are required for a waiver generally need to be more robust.

7.2. OVERVIEW OF THE EOGRTS TEST GUIDELINE

The test is designed to evaluate the effects of chemicals on pre- and postnatal development; the integrity of the male and female reproductive systems; and systemic toxicity in males, pregnant and lactating females, and young and adult offspring [OECD 443]. An overview of the test design is represented schematically in Figure 7.1B.

The test substance is administered to groups of sexually-mature male and female animals for a minimum of two weeks prior to mating, although up to ten weeks may be preferred under some regulatory frameworks; during the mating period, of maximum two weeks' duration; and throughout gestation and lactation until weaning of the offspring (females); or at least until weaning of the F_1 offspring, and for a minimum of ten weeks (males). Selected offspring are further administered the test substance until the time of scheduled necropsy. Litter size may be adjusted at postnatal day (PND) 4, and again at weaning once F_1 animals have been assigned to their respective cohorts.

Clinical observations and pathology examinations are undertaken on all animals to evaluate signs of toxicity, with particular emphasis on the integrity of the male and female reproductive systems, and the growth, development, and function of the offspring. At weaning, offspring may be assigned to cohorts to further evaluate the effect of the test substance upon specific developmental and reproductive endpoints.

Litter size may be adjusted according to the normal litter size of the strain being used, so that there are eight to ten F_1 offspring at the time of weaning [OECD 443, OECD GD 151]. They are assigned to cohorts in order to examine the effects of the test substance upon specific aspects of development: cohort 1, reproduction; cohort 2, DNT; cohort 3, DIT:

- Cohort 1A. One male and one female from each litter are assigned to evaluate effects upon the reproductive system and general systemic toxicity.
- Cohort 1B. One male and one female from each litter are assigned to evaluate reproduction and fertility of the F_1 generation, if triggered, or otherwise to confirm equivocal effects upon the reproductive system in cohort 1A.
- Cohort 2A. One male or one female from each litter is assigned for neurobehavioral testing and subsequent neurohistopathology assessment as adults.
- Cohort 2B. One male or one female from each litter is assigned for neurohistopathology assessment at weaning.
- Cohort 3. One male or one female from each litter is assigned to evaluate T-cell dependant antibody response at eight weeks of age.
- Surplus. Remaining unselected pups are subjected to gross pathology examination and used for T4 and TSH on PND 22.

All F_1 animals from cohorts 1, 2A, and 3 are evaluated for age at which sexual maturation is attained and for external abnormalities of the reproductive tract organs.

7.3. USE OF AVAILABLE DATA IN PLANNING THE EOGRTS

The EOGRTS is a resource-intensive and complex study, and it will not be performed in the absence of data that at least allow adequate dose selection [OECD 443]. Attention herein is given to information that is routinely available within the tiered and integrated testing strategies that are required under the standard regulatory frameworks for industrial and agricultural chemicals. As advances are made in alternative screening methods, additional tests may become a routine part of the standard testing strategies, thereby providing additional information to the decision-making process.

This guidance begins with the assumption that the EOGRTS will generally be performed without the modules for DNT (cohort 2) and DIT (cohort 3), and that a breeding to produce a second generation (cohort 1B) in not necessary. Prior to the conduct of the EOGRTS, relevant data for the identification of triggers could be available from one or more of the following sources[1]:

- In the absence of any directly relevant data for the substance under consideration, structure-activity relationships or reference to class chemicals for which relevant data are available may be useful. Given the complexity of the EOGRTS some data for the substance will likely be available, from dose range-finding tests, for example; but data for specific triggers may not. Note that lack of identified SAR concerns should not be construed as a waiver.
- Acute toxicity data in rodents [OECD 420, OECD 423 OECD 425]. Acute toxicity tests may be of some use with respect to the presence of effects upon the nervous system. Compared to repeated-dose toxicity study designs exposures or doses in acute tests are high; are administered once; and oral doses are given by gavage resulting in high peak plasma concentrations. Furthermore, routinely-evaluated endpoints are limited to in-life clinical observations and gross pathology. Data from repeated-dose toxicity studies may be more relevant.
- Repeated dose toxicity in rodents [OECD 408, OECD 407, OECD 413, OECD 422]. In-life observations during the conduct of 28-day and 90-day repeated dose toxicity studies may reveal potential effects upon the central and peripheral nervous systems, endocrine system, and immune system.
- Reproductive/developmental toxicity screening studies or endocrine disrupter screening studies in rodents [OECD 421, OECD 422 OECD 440, OECD 441]. In addition to signs of adverse reaction to treatment relevant to the selection of doses for the conduct of the EOGRTS [OECD 443], screening studies may provide initial or supporting evidence for adverse effects upon fertility or development that could require further evaluation, or supporting evidence for equivocal findings in cohort 1A. These study designs are, however, limited in terms of statistical power and/or mechanistic scope when compared to the EOGRTS; negative findings on their own should not be construed as conclusive evidence of no concern and thus a waiver against further evaluation.
- Prenatal developmental toxicity [OECD 414]. The prenatal developmental toxicity study design examines a specific subset of the reproductive cycle, and may foretell the manifestation of morphological or functional deficits during postnatal development.

1 Where specific OECD Test Guidelines are noted, equivalent regulatory test guidelines also apply.

7.4. TECHNICAL TRIGGERS AND WAIVERS FOR THE SPECIFIC COHORTS

Triggers fall into one of two broad categories: the identification of effects that predict potential or concern for toxicity in the offspring; and effects in adults that raise concerns for the developing offspring as a sensitive subpopulation. Cohort 1A is integral to the EOGRTS, and the application of triggers applies only to the mating of F_1 animals in cohort 1B to produce a second generation. Given the resource intensiveness and logistical complexity of the EOGRTS, triggers for cohorts 2 and 3 are only envisaged from prior information that is available during the design phase of the study, and not from observations made during the conduct of the study itself (Table 7.1).

When evaluating endpoints as triggers for a study their *strength of association* and *specificity* need to be considered. *Strength of association* indicates confidence in a true causal relationship between effect and treatment. Effect frequency, magnitude relative to background, and treatment correlation (including time course and dose response) are primary contributors to strength of association. *Specificity* is the degree to which an effect is qualitatively related to treatment with the test item, whether it's a primary effect of treatment or arises secondarily to another response. Together they determine the confidence that the response is a treatment-related effect of regulatory concern, as opposed to a chance or secondary response. Effects that are clearly related to treatment (either in the EOGRTS or from previous studies) may not require confirmation if they are sufficient to conclude on hazard identification (classification) and risk assessment, such effects effectively become waivers; those that are equivocally related to treatment (for example achieving only borderline statistical significance, unclear dose-relatedness, or minor deviation from historical control) may require confirmation.

Reproduction is a highly complex and integrated function, and sometimes changes in measured endpoints reflect non-specific responses to stress or biological variability ([Moore 2013]. When identifying parameter changes as triggers for further investigation, care must be taken to ensure that the effect is a direct response to treatment and is not a secondary or non-specific response. Proper statistical analysis is critical when making this determination, particularly when endpoints are interdependent, but statistical significance should not be the sole determinant. OECD guidance on the evaluation of repeated-dose and reproductive toxicity data is available, and following this guidance is strongly recommended [OECD GD 32, OECD GD 35, OECD GD 43]. The final decision on trigger selection should be made using expert judgement and must be fully justified in the final report.

Table 7.1. Triggers and waivers for additional cohorts within the EOGRTS.
Effects that are clearly related to treatment, and for which decisions on hazard identification and risk assessment can be made, as well as effects that are clearly a secondary response to systemic toxicity may not necessarily result in further investigation. The final decision should be based on the weight of all available evidence.

Cohort	Internal triggers (+) and waivers (–)		External triggers
	F_0	F_1	
1B (F_2 generation)	+ ↓ Matings + ↓ Fertility (implantations, pregnancy rate) + ↑ Gestation length – ↓ Reproductive organ weights, pathology – Sperm parameters	+ ↓ Litter size + ↓ Litter/pup weight + Anogenital distance + Nipple/areola retention + Age at sexual maturation + Oestrus cycle	+ Evidence of an endocrine disrupting mode of action
2A & 2B (DNT)	None (impractical)	None (impractical)	+ SAR or chemical class (neurotoxicity) + Evidence of mechanistically-relevant endocrine disrupting activity + Evidence of a relevant neurotoxic mode of action + Functional or morphological neurotoxicity in adults + ↓ Surface righting reflex + ↓ Thyroid weight and pathology
3 (DIT)	None (impractical)	None (impractical)	+ SAR or chemical class (immunotoxicity) + Evidence of mechanistically-relevant endocrine disrupting activity + Altered weight or pathology of lymphoid organs + Altered cellularity of bone marrow, spleen, thymus, or lymph nodes + Altered differential white cell counts + Functional immunotoxicity in adults

7.4.1. Triggers for cohort 1B (breeding to produce the F_2 generation)

ACSA proposed that second breeding to be triggered by adverse effects in the parental animals or F_1 progeny, specifically impaired fertility or fecundity of the parental generation; abnormal sexual development of the F_1 pups; or postnatal mortality or toxicity to the F_1 pups prior to weaning. The OECD published guidance on triggers for the second generation breeding to be applied to studies conducted under the Canadian Pest Management Regulatory Agency and the United States Environmental Protection Agency's Office of Pesticide Programs. The triggers largely reflect those originally proposed by ACSA [Cooper 2006], specifically effects on F_0 fertility; F_1 oestrus cycle; litter size; F_1 developmental landmarks; postnatal mortality; pup malformations; decreased live birth index; and decreased pup body weight [OECD GD 117]. ACSA had recognised that, due to the complexity and interrelatedness of the evaluations, it was likely that any critical or equivocal effect on reproduction would be detected and would trigger breeding of the F_1 generation [Cooper 2006]. Indeed, a review of nine two-generation reproductive toxicity studies indicated that the triggers proposed by ACSA would have resulted in the second generation breeding in five out of six cases in which reproductive effects were observed [Beekhuijzen 2009].

The application of triggers should be exercised in recognition of the evidence that the second breeding adds little or no impact to the overall hazard and risk assessment [Janer 2007; Piersma 2011; Rorije 2011], and considering the feasibility of their application under specific circumstances [Fegert 2012]. Unless dictated by specific regulatory frameworks the second breeding should be triggered only to confirm otherwise equivocal findings. Effects that are clearly related to treatment should not require confirmation if they are sufficient to conclude on hazard identification (classification) and risk assessment. Likewise, the absence of any effect on reproduction toxicity would not require further testing. Thus the unequivocal presence or absence of effects are effectively waivers for cohort 1B.

With regards to the specificity of effects, the following considerations should be taken into account:

- The most sensitive endpoints to determine effects upon fertility in rats are organ weights, sperm parameters, and histopathology, rather than litter size [Blazak 1989; Moore 2013]. The assessment of these endpoints in the F_1 generation is a requirement of the test guideline, so if reduced litter size or implantations is observed in the presence of significant, biologically-relevant changes in these other endpoints then a second breeding should not be required because it would not alter the overall assessment. Since these endpoints are interrelated, the assessment should take into account all available data. The assessment of organ weights should take into account potential influence of body weight; it may be preferable to apply analysis of covariance (ANCOVA) rather than express organ weights relative to body weight [Shirley 1977].
- Litter size is influenced by the number of ova released prior to conception, which is reflected by the number of *corpora lutea* that remain to support conception. Reduced *corpora lutea* counts have been related to low body weight prior to mating [Chapin 1993], and thus care

should be taken to ensure that reductions in litter size cannot be attributed as a secondary response to effects upon maternal weight. It is not feasible to count *corpora lutea* in the context of the EOGRTS, but data from other relevant studies may prove useful.[2]

- Offspring body weight at birth may be influenced by maternal body weight, as a reflection of nutritional status; litter size; and proportion of males in the litter [Carney 2004; Griggio 1997; OECD GD 43; Romero 1992]. Therefore, identification of a specific treatment-related effect upon offspring weight should also consider these factors, preferably using an *ad hoc* method, such as ANCOVA.
- Anogenital distance is a marker not only of sexual differentiation but also of general growth, and therefore it should be analysed by ANCOVA using pup body weight as covariate; otherwise it should be normalised to the cube root of pup body weight [Gallavan 1999].
- A delay or acceleration in the age at which sexual maturation is attained (opening of the vagina, separation of the prepuce from the glans penis) may trigger concern. Again, these landmarks are influenced by overall growth of the animal, and weight at attainment should be taken into consideration [Melching-Kollmuß 2014].
- When considering endocrine activity data, greater weight should be given to the results of relevant apical studies *in vivo*, which more accurately reflect the dynamic situation of the EOGRTS, rather than mechanistic or screening studies *in vitro*, which do not take into account the dynamics associated with toxicokinetics and the interrelatedness of the reproductive cycle.
- Malformations in the F_1 generation should not be a trigger for the 1B cohort. Confirmation of malformations is better achieved through the conduct of a prenatal developmental toxicity study. If a prenatal developmental study is already available, the findings should be considered *in toto* and a decision then made as to whether further confirmation through a second breeding would be meaningful.

7.4.2. Triggers for cohort 2 (developmental neurotoxicity)

Triggers for the conduct of a stand-alone DNT test, such as OECD TG 426 [OECD 426], have been applied to testing of agricultural chemicals. A workshop held to discuss triggers, among other aspects of DNT testing under the US Environmental Protection Agency, outlined several that might arise from available data on a compound, along with levels of concern based on biological effect [Levine, 1990]. The biologically-related triggers were substances that were CNS or behavioural teratogens; adult neuropathic and neuroactive toxicants; hormonally active compounds; and selective developmental toxicants without CNS effects. The last of these is less relevant as a trigger for a DNT module within the EOGRTS because such substances may be identified within cohort 1, and are better characterised by an appropriately designed prenatal developmental toxicity study.
In repeated dose studies, particularly those required for regulatory purposes and conducted according to OECD guidelines, general clinical examinations are performed on a daily basis as well as

2 Note that while *corpora lutea* counts are undertaken in the prenatal developmental toxicity study [OECD 414], treatment is usually initiated some days after ovulation. Furthermore, while earlier versions of the reproductive toxicity screening studies allowed assessment of *corpora lutea* under meaningful treatment conditions, the current test guidelines do not [OECD 421, OECD 422]. Useful information may, however, be derived from reproductive toxicity screening studies conforming to earlier guidelines.

weekly detailed clinical examinations. Examination of clinical signs include potential occurrence of secretions and excretions and autonomic activity (e.g. lacrimation, piloerection, pupil size, unusual respiratory pattern). Detailed clinical observations should be performed in an open arena to optimise the detection of anomalies at the time of the evaluation. Changes in gait, posture, and response to handling as well as the presence of clonic or tonic movements, stereotypies (e.g. excessive grooming, repetitive circling), or bizarre behaviour (e.g. self-mutilation, walking backwards) are also recorded. In addition, sensory reactivity to stimuli of different types (e.g. auditory, visual and proprioceptive stimuli), assessment of grip strength and motor activity assessment should be conducted; a pre-test observation battery is recommended to establish the baseline. Pathological examination of brain (representative regions including cerebrum, cerebellum and medulla/pons), spinal cord (at three levels: cervical, mid-thoracic and lumbar), peripheral nerve (sciatic or tibial) preferably in close proximity to the muscle is also performed. Changes in these parameters could reflect neurotoxicity affecting the central as well as peripheral nervous system and may form the basis to trigger the conduct of cohort 2. In addition, knowledge of the mode of action of the chemical (e.g. cholinesterase inhibition, interaction with receptors involved in neurotransmission) should be used as supporting evidence.

Oestrogens play a role in CNS development [Ferguson 2000], and neurotoxicity has been expressed by neuroendocrine disrupters *in vivo* [Waye 2011]. 'Neuroendocrine disrupters' are broadly defined as substances that interfere with the metabolism or action of "neuropeptides, neurotransmitters, or neurohormones" [Waye 2011]. Of these, probably only the last, and possibly the first, fall into the more broadly accepted concept of 'endocrine disrupter'. Nevertheless, evidence of endocrine disrupting activity in these systems might be a cause for further evaluation.

With regards to the specificity of effects, surface righting reflex is a measure of development and coordination of the vestibular and neuromuscular systems. Delayed acquisition of the reflex may reflect delays in motor development or function, but can also be affected by general growth of the offspring. Consequently, care should be exercised not only to take into account the litter as the unit of analysis, but also those factors that influence offspring body weight described above—body weight, litter size, and proportion of males in the litter—should be considered as possible covariates.

7.4.3. Triggers for cohort 3 (developmental immunotoxicity)

The issues of children's susceptibility to immunotoxicants and DIT testing have been the subject of much recent significant research, reviews, and workshop discussions [e.g. workshops reported by Boverhoff 2014; Holsapple 2012; Holsapple 2005; Piersma 2012]. Of most significance for the determination of triggers for DIT is the apparent greater sensitivity of the developing immune system compared to the mature system present in adults and the relative sensitivity of functional parameters compared to necropsy findings [Luebke 2006; Smialowicz 1988; Tonk 2010, Tonk 2011, Tonk 2011, Tonk 2012, Tonk 2013, Tonk 2013; WHO 2012].

Some assessment of the 'steady state' immune system competence following exposure to a test substance is undertaken as a core part of the EOGRTS. Splenic cellularity and haematology are evaluated in cohort 1A; the spleen and thymus from all F_1 adults are weighed; and spleen, thymus, and samples of bone marrow from all F_1 adults are preserved for possible histopathology. But these do not represent a functional response to antigen exposure. Functional response, which integrates several aspects of the immune system, is determined in cohort 3 with a T-cell dependant antibody response (TDAR) assay. The TDAR assay measures the primary antibody response, and is useful in detecting chemicals that suppress adaptive immunity. In the EOGRTS cohort 3 a TDAR assay occurs around postnatal day 56 [OECD 443].

While the developing immune system is recognised as more sensitive to toxic insult than that of the adult [Luebke 2006; Smialowicz 1988; Tonk 2012], assessments to compare sensitivity have largely deviated from the TDAR assessment in OECD TG 443, particularly with reference to the age at challenge and the parameters that were assessed, and therefore application to the concept of triggers within the requirements of the test guideline is weakened. Luebke *et al*. [Luebke 2006] reviewed the immunotoxicity data for five substances—diethylstilboestrol; diazepam; lead; 2,3,7,8-tetrachlorodibenzo-*p*-dioxin; and tributyltin oxide—demonstrating that the developing immune system was more sensitive in terms of the doses that elicited an adverse response or persistence of response; although in most cases the comparison was complicated by differences in dose selection, endpoints reported, species, or strain. In the case of tributyltin oxide, however, the range of effective doses in immature and mature animals was similar, and thymus atrophy was one of the most sensitive endpoints, although effects were more persistent in immature animals [Luebke 2006]. The lymphocyte proliferation response to T- and B-cell mitogens was depressed and persisted to a greater extent in male Fischer 344 rats that were administered dioctyltin dichloride as neonates compared to adults. Thymus weight was depressed to a similar extent in both age groups at the earliest evaluation time points following cessation of exposure, but was unaffected at later time points when the lymphocyte proliferation response was still depressed, and NK cell activity was unaffected [Smialowicz 1988]. Juvenile rats also appeared more sensitive to immunotoxicity induced by di-(2-ethylhexyl) phthalate in terms of changes in white blood cell counts, and keyhole limpet haemocyanin (KLH)-stimulated lymphocyte proliferation and cytokine response *ex vivo*, but there was no difference between the age groups in IgG response and no change in IgM [Tonk 2012].

Tonk and co-workers have evaluated the relative sensitivity of several measures of immune status following exposure of juvenile rats to di-(2-ethylhexyl) phthalate [Tonk 2012]; dioctyltin dichloride [Tonk 2011, Tonk 2011]; ethanol [Tonk 2013, Tonk 2013]; 4-methyl anisole [Tonk 2015]; and methyl mercury [Tonk 2010]. Immune status was variously assessed by weight of lymphoid organs; haematology; spleen and thymus cellularity; mitogen-stimulated cell proliferation; KLH-stimulated primary and secondary serum IgG and IgM titres; and KLH-stimulated lymphoproliferation and cytokine release in cultured splenic lymphocytes following an additional 96-hour *ex vivo* exposure to the antigen. Antigen challenge and response was variously conducted at around four, eight,

or ten weeks of age. In most cases, IgM was not elevated by treatment, and while the relative sensitivity of other endpoints varied they were generally of a similar order of magnitude. Of note, the sensitivity of lymphoid organ weights and spleen cellularity, which along with IgM are the only parameters that were investigated that are required as part of the EOGRTS [OECD 443], did not differ greatly to that of IgG, KLH-mediated proliferation, and cytokine release. IgG and cytokine release may be incorporated into follow-up investigations to the EOGRTS if necessary [Boverhoff 2014; OECD GD 151].

Consequently, while there are differences in the sensitivity of the immature and mature immune systems to toxic insult, particularly with respect to persistence, and some functional parameters may be more sensitive, there is no evidence that a combined evaluation of lymphoid organ weights, histopathology, and cellularity is less sensitive than plaque-forming cells, one of the additional evaluations required in the DIT module (Table 7.2). If cohort 3 is not triggered due to lack of such findings in previous studies, but then effects on lymphoid organs or cellularity are identified in the EOGRTS F_1 adults, they would trigger further evaluation in a separate DIT study. Further information on the responsiveness of the endpoints that are routinely evaluated in repeated dose toxicity studies and the EOGRTS may be helpful in clarifying the sensitivity of the available triggers.

Table 7.2. Sensitivity and specificity of immunotoxicity endpoints.
An analysis of the sensitivity, specificity, and concordance of organ endpoints that are evaluated in repeated dose toxicity studies with immunotoxicology classification for 51 substances tested in adult female B6C3F1 or C57BL/6 mice. Immunotoxicology classification was assigned on the basis of performance in specific functional assays [Luster 1992].

Endpoint or combination of endpoints	Number of chemicals	Sensitivity	Specificity	Concordance
Organ endpoints				
Thymus weight relative to body weight	40	0.54	1.00	0.68
Spleen weight relative to body weight	41	0.52	0.83	0.61
Spleen cellularity	36	0.38	0.92	0.56
Haematology				
Leukocytes	30	0.20	0.90	0.43
Combination effects				
Thymus *and* spleen weights relative to body weight	40	0.36	0.83	0.50
At least one organ endpoint	45	0.77	1.00	0.84
Any two of three organ endpoints	40	0.46	0.92	0.60
All three organ endpoints	32	0.13	0.78	0.31
At least one organ endpoint or leukocytes	46	0.81	1.00	0.87
Functional observations				
Plaque-forming cell (PFC) response	45	0.54	1.00	0.64

Oestrogens play a role in the development and control of immune function [Cunningham, 2011; Soucy 2005; Verthelyi, 2006], and immunotoxicity has been expressed by oestrogen agonists and antagonists [Kalland 1979; Luebke 2006]. Consequently, evidence of endocrine disrupting activity *in vivo*, especially oestrogenic or anti-oestrogenic, might be a cause for further evaluation.

In repeated dose studies—particularly those required for regulatory purposes and performed according to OECD guidelines—blood samples are taken before necropsy; total and differential leucocyte counts might provide indications of an immunotoxicological mode of action. At necropsy, the spleen and thymus are weighed, fixed, and examined for histopathological changes. Histopathological examinations are also done for mesenteric and axillary lymph nodes, Peyer's patches and bone marrow.

Indications of immunosuppression in standard rat toxicity studies [OECD 422, OECD 408, OECD GD 32, OECD 407, OECD 413] include haematology, clinical chemistry, and anatomical pathology. Relevant haematological endpoints are leukopenia, if not associated with severe body weight decrease, severe kidney damage, or multi-morbid status; neutropenia, if not associated with anaemia, severe kidney damage, liver cirrhosis, or hypothyroidism; lymphopenia, if not associated with acute inflammation, damage to lymph nodes, chronic renal failure, severe stress, hyperadrenocorticism, or severe heart failure; and eosinopenia, if not associated with stress, hyperadrenocorticism, or acute inflammation. The determination of hypoglobulinaemia as part of the clinical chemistry battery may indicate immunosuppression if noted in the absence of other causes such as severe liver or kidney damage, haemorrhagic or haemolytic anaemia, over-hydration, or dysproteinaemia affecting transport globulins. The non-immunosuppressive causes for alterations in these haematological and clinical chemistry parameters can be excluded by changed standard clinical pathology parameters like liver enzymes, urea, creatinine, red blood cell parameters and by histopathology investigations.

Decreased organ weights of thymus, spleen, lymph nodes in combination with histopathology findings, like atrophy and cell depletion in spleen, thymus, lymph node, and bone marrow at lower doses than those eliciting systemic toxicity may be an indicator of immunotoxicity. Isolated weight alterations of the mentioned organs without histopathology findings are rather more an expression of systemic toxicity than immunotoxicity. Atrophy of the thymus in subchronic studies but not in chronic studies may be an indicator of immunotoxicity [Kuper 2000; US EPA 2013].

Differentiation of immunosuppression versus stress

It is crucial to differentiate between a primary immunosuppressive effect and subacute stress in repeated dose studies. A primary immunosuppression can be expected if the effects in lymphoid organs (atrophy, hypocellularity, leukopenia etc.) are found in the absence of other adverse effects. Stress as a cause of the alterations in immune system parameters may be confirmed if the effects tend not to be dose-dependent and if occurring concomitantly with either decreased food consumption and body weight, decreased thymus weight in the absence any other altered lymphoid organs, hypertrophy of the zona fasciculata of the adrenal cortex, or neutrophilia combined with eosinopenia [Everds 2013].

Additional immunotoxicity parameters in standard toxicity studies

If QSAR analysis, toxicokinetic data, or previous study results reveal a potential for an immunosuppressive potential of the compound, additional parameters can be included in the subsequent regulatory studies (e.g. subchronic repeated dose toxicity study) *without increasing the animal numbers*. Regarding clinical pathology, the following parameters can be included in rodent studies: serum electrophoresis, to differentiate between a deficiency of transport globulins and immunoglobulins; immunoglobulins, direct immunologic measurement of immunoglobulin fractions in serum (IgG, IgM, eventually IgA and IgE); and cytokines. The direct measurement of blood cytokine levels is a poor indicator of functional suppression of the immune system, because normal serum levels are very low and a decrease cannot be measured [House, 1999]. However, if there is an indication of, or concern for, immunosuppression, blood and/or spleen cells isolated from rats in repeated dose studies can be used to measure cytokine production [Ai 2013, Ai 2014; Barten 2006]. For comparison reasons a standard protocol for these *ex vivo* assays will be necessary [Dietert 2007].

While skin sensitisation data may be available prior to an EOGRTS, it is unlikely to be useful informing a decision to trigger cohort 3. The TDAR assays required as part of OECD 433 detect immune suppression, but not immune enhancement or immune responses directed toward chemicals. There are many agents that can affect TDAR assays, and there is no single Adverse Outcome Pathway that encompasses all of them. In contrast, skin sensitisation assays assess the potential for the adaptive immune system to recognise and attack chemically-modified proteins. There is a single OECD-accepted Adverse Outcome Pathway describing how the process unfolds for a multitude of chemicals [OECD No. 168, 2014]. In essence, a finding of sensitisation is a finding that a memory immune response directed toward chemically-modified proteins can develop. However, this does not alter the ability to generate memory responses toward other targets of immunity (e.g., bacteria, viruses, or tumours). Development of a memory response directed toward a specific chemically-modified protein, the domain of skin sensitisation assays, generally has little implication on the global function of the adaptive immune system, the domain of TDAR. A change in one has no anticipated biological relationship to the other, and consequently evidence of skin sensitisation should not be construed as a trigger for cohort 3.

7.5. ADDITIONAL TRIGGERS

Further to the triggers that have been suggested herein, additional triggers may be stipulated within specific regulatory frameworks. These may reflect regulatory policy towards wider registration requirements and an overall testing strategy, and may not be strictly related to the biological relationships described above. The practitioner is guided to consult the specific regulatory requirements in such cases.

Acknowledgements

This work was conducted as a task force under the auspices of ECETOC. NPM acted as a paid consultant to ECETOC in the conduct of this work, and drafted the manuscript; the other authors were engaged in the course of their normal employment. The authors are grateful to Dr. Marie-Louise Meisters (DuPont) for critical review of the manuscript. Other members of the task force were Joseph Lewis (DuPont), and Dr. Sue Marty (Dow Chemical).

Chapter 8

From cutting edge to guideline: a first step in harmonization of the zebrafish embryotoxicity test (ZET) by describing the most optimal test conditions and morphology scoring system

Manon Beekhuijzen[a], Coco de Koning[b], Maria-Eugenia Flores-Guillén[c], Selinda de Vries-Buitenweg[a], Marysia Tobor-Kaplon[a], Beppy van de Waart[a], Harry Emmen[a]

[a] WIL Research Europe B.V., Hambakenwetering 7, 5231 DD, 's-Hertogenbosch, The Netherlands.

[b] Toxicology, Radboud University Medical Centre, Nijmegen, The Netherlands.

[c] Toxicology and Environmental Health, Utrecht University, Utrecht, The Netherlands.

Published in: Reproductive Toxicology. 2015 Aug 15;56:64-76. doi: 10.1016/j.reprotox.2015.06.050. Epub 2015 Jun 22.

ABSTRACT

In the last couple of years, the interest in the zebrafish embryotoxicity test (ZET) for use in developmental toxicity assessment has been growing exponentially. This is also evident from the recent proposal for updating the ICHS5 guideline. The methodology of the ZET used by the different groups varies greatly. To further evaluate its successfulness and to take the ZET to the next level, harmonization of procedures is crucial. In the present study, based on literature and empirical data, the most optimal study design regarding temperature, test chamber, exposure period, presence of chorion, solvent use, exposure method, choice of concentrations, and teratogenic classification is proposed. Furthermore, our morphology scoring system is reported in detail as protocol to further enhance study design harmonization.

Keywords
Zebrafish, zebrafish embryotoxicity test (ZET), teratogenicity, morphology scoring system, harmonization, 3R's

8.1. INTRODUCTION

The zebrafish (ZF) has emerged as a promising alternative animal model to screen for teratogenicity in preclinical developmental toxicity assessment, because it is cheaper and faster, when compared to tests with mammals, and complies with the ethical framework of the 3R's. In addition, this model enables continuous developmental monitoring of a whole organism from early embryogenesis until hatching and beyond. It is therefore, that the ZF embryotoxicity test (ZET) was recently proposed as update in the ICH S5 guideline [ICH S5(R3) Draft January 2014]. The draft concept paper mentions "a combination of *in vitro* assays, *in silico* prediction, and ZF assays is advocated rather than relying on the use of one stand-alone method".

Since 2009, a consortium of six biopharmaceutical companies are working hard to develop a harmonized ZET [Brannen 2010, Panzica-Kelly 2010, Gustafson 2012], however still technical differences remain and are the primary contributor to inter-laboratory differences in predictivity [Ball 2014]. However, until now, no global harmonized experimental setup has been established. This is crucial for further incorporation of the ZET worldwide and eventually might lead to implementation of the ZET for registration purposes.

For starting this harmonization process, one of the most meaningful things to move the ZET forward would be the publication and sharing of as many predictive models as possible. Key features will be: 1) which developmental endpoints are measured, 2) when they are measured, 3) whether a measure of cytotoxicity is included in the predictive model, and 4) whether it is useful to know how much compound actually gets into the embryos [Chapin 2008]. Moreover, we would encourage also to publish unsuccessful data. This will enrich further development and improvement of the ZET methods.

The aims of this paper are to evaluate the ZET experimental setups in literature and describe the study design that would yield high predictability rates, yet be pragmatic, and to further advance harmonization of the ZET study design.

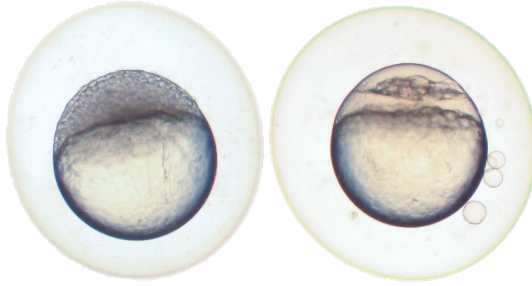
To achieve these aims, a thorough analysis of published peer-reviewed manuscripts on ZF embryotoxicity tests was done to compare and select all those test conditions that are considered to impact ZET outcome. Test conditions such as type of test chamber, exposure length, dechorionation, use of solvent, exposure method, and teratogenicity classification approach have been reported to play a crucial role in the correct classification of compounds when using the ZET, and will be discussed in the following paragraphs. As a basis, the in 2013 adopted and fully validated OECD Guideline for the fish embryo acute toxicity (FET) test (i.e. OECD 236) [OECD 236] was used; modifications and additions were made based on this literature review. Additionally, the impact of temperature (26 or 28°C) is presented through preliminary data to aid in this thorough evaluation. Furthermore, the proposed morphology scoring system is reported in detail, as protocol, to enhance study design harmonization.

8.2. SELECTION OF THE MOST OPTIMAL TEST CONDITIONS FOR THE ZET

In the next paragraphs, the following test conditions are discussed in detail: selection of eggs, start of exposure, test temperature, dechorionation, type of test chamber, exposure length, choice of concentrations, exposure method, and use of solvent.

8.2.1. Selection and number of eggs

To improve survival of ZF embryos, the ZET should be started only with fertilized eggs showing a round and transparent chorion and presenting no signs of coagulation. Early signs of coagulation include formation of bubbles in the egg (especially around the animal pole) and asymmetry of the animal pole or yolk. These features are illustrated in Figure 8.1.



A: normal ZF egg at 4 hpf

B: ZF egg at 4 hpf showing signs of coagulation; asymmetric animal pole and bubbles around the embryo.

Figure 8.1. ZF egg showing signs of coagulation at 4 hours post fertilization (hpf) (B), and normal 4 hpf ZF embryo (A). Bubbles and asymmetry can be observed in the embryo (B), which are absent in the normal embryo (A).

From the literature search it was clear that a wide range in number of eggs are being used. These numbers vary from 2 embryos per concentration [Padilla 2012] to 48 embryos per concentration [Ali 2011]. In the OECD validation study to assess intra- and inter-laboratory reproducibility of the FET test, evaluation of the effect of the group size confirmed that the use of 20 embryos per concentration was preferable [Busquet 2014]. They tested 20 chemicals at 5 concentrations in 3 independent runs in at least 3 laboratories. Specifically, the more shallow the concentration-response curve, the greater the benefit of 20 embryos per group versus the smaller sizes evaluated [Busquet 2014]. As these numbers are needed in the FET for evaluation of the LD50 (concentration with 50% mortality), we would recommend to use at least 30 embryos per concentration in the ZET for assessment of teratogenicity as more variation can been seen with this endpoint compared to mortality.

8.2.2. Start of exposure

In literature, the start time of exposure varies in high degree from study to study. Some research groups start exposure at ≤2 hours post fertilization (hpf) [Fraysse 2006, Hermsen 2011, Selderslaghs 2012] while others start exposure at 6-8 hpf [Padilla 2012, Huang 2010, Truong 2011 or even at 24 hpf [Ali 2011].

The initial goal of the ZET as a complement for developmental toxicity studies suggests exposure of ZF embryos should start at the same embryonic stage in which maternal exposure starts in these studies. In rat and rabbit developmental toxicity studies (e.g. ICH S5(R2) and OECD 414) [ICH S5 (R2) 2005, OECD 414] maternal exposure starts after implantation is completed, which is at Day 6 post-coitum also known as the gastrulation period. The gastrulation stage in the ZF embryo is equivalent to 4-5½ hpf [Kimmel 1995]. Therefore, the best time to start exposure is at 4 hpf, covering the early gastrulation stage. This will facilitate the extrapolation of results from those obtained from (non)rodent developmental toxicity studies as exposure would cover the same development stages.

8.2.3. Temperature

Based on our literature review, current ZF embryotoxicity tests are performed at 26 to 28.5°C. A temperature of 26 or 26.5°C is used by several research groups [Padilla 2012, Fraysse 2006, Hermsen 2011, Busquet 2008, Hallare 2004, Weigt 2011], and is also indicated by the FET guideline (OECD 236) [OECD 236]. A temperature of 28 or 28.5°C is used by several other research groups [Brannen 2010, Selderslaghs 2012, Huang 2010, Truong 2011, van den Bulck 2011]. The decision on optimal temperature is one of the most critical, as temperature has a direct impact on the development of the ZF embryo and its capacity to cover its developmental milestones, as development of ZF slows at lower temperatures [Beasly 2012]. In addition to its effect on ZF development rate, a higher temperature could also induce evaporation which could have a direct impact on the internal well concentration, and might therefore have a major impact on the study outcome for volatile test compounds.

In two preliminary experiments, the effect of temperature (26 and 28°C) on teratogenicity scoring in the ZET was evaluated using the positive controls 3,4-Dichloroanaline (3,4-DCA; CAS 95-76-1) at 1.5 mg/L and 3 mg/L , and Dimethyl sulfoxide (DMSO) (CAS 67-68-5) at 0.5 % and 1 % in ISO-medium (Table 8.1). Per concentration 10 embryos were exposed.

Table 8.1: Fraction of ZF larvae showing teratogenic effects after exposure to 3,4-DCA or DMSO at 26 or 28˚C.

	26 ˚C	**28 ˚C**
1.5 mg/L 3,4-DCA (experiment 1)	6/10 (60%)	1/10 (10%)
1.5 mg/L 3,4-DCA (experiment 2)	3/7 (43%)	0/6 (0%)
3 mg/L 3,4-DCA (experiment 1)	8/8 (100%)	6/6 (100%)
3 mg/L 3,4-DCA (experiment 2)	7/7 (100%)	4/4 (100%)
0.5 % DMSO	2/10 (20%)	2/9 (22%)
1 % DMSO	10/10 (100%)	3/7 (43%)

As apparent from Table 8.1, ZF larvae exposed to 3,4-DCA, a positive control for teratogenicity, and cultured at 26 ˚C show more teratogenic effects, already at lower concentrations of the compounds tested than ZF larvae cultured at 28 ˚C. Also, survival of ZF larvae was higher at 26 ˚C. In addition, exposure to 1% DMSO at 26 ˚C showed more teratogenic effects than at 28 ˚C. From these preliminary results, it appears that 26°C provides a worst-case-scenario.

We suspect that the difference between incubation at 26°C and 28°C is a result of quicker ZF development with increasing temperature. Therefore, it can be postulated that teratogenic effects at 28°C are unnoticed due to a limited time window for development assessment, or that embryos incubated at lower temperatures might remain longer in a specific developmental stage in which certain developmental endpoints are more sensitive to teratogenic effects.

This emphasizes that daily assessments are of great added value as information on teratogenicity could be missed when only evaluating on the last day of the test. The advantage for assessments on a daily basis is that it provides crucial information on the development of certain findings over time and might elucidate dose response relationships. Moreover, any morphological abnormalities early in development will be scored, which would be missed when only evaluating on the last day as at that time the malformed embryo/larva would be coagulated, and it will then be scored as mortality instead of malformation. It is of utmost importance to make this distinction.

Even though these limited studies strongly suggest that incubation at 26°C enhances predictability of the ZET, more research is needed to provide definitive conclusions on the most suitable temperature for the ZET.

8.2.4. Chorion

The chorion as a barrier for compound uptake is nowadays a matter of many debates and discussions. This envelope covers the embryo up to 48-72 hpf when cultured at 26 to 28°C; it possesses small pore canals of 500-700 nm in diameter [Lee 2007]. Despite repeated speculations, it could not be clarified unequivocally whether the chorion represents an effective barrier and, thus, effectively protects the embryo from exposure to distinct chemicals. The pores have been reported to potentially restrict the uptake of compounds depending on their size, as could be shown for flu-

orescent dextrans exceeding 3 kDa [Creton 2004]. There is limited evidence that the permeability of the chorion changes during the embryonic development and that, after hardening, the chorion is less permeable [Gellert 2001]. It then functions as a barrier for even smaller molecules for the ZF embryo [Braunbeck 2005]. Gustafson et al. [Gustafson 2012] concluded from their optimization studies that dechorionation (either manually or by protease treatment) added complexity and assay set up time with only small improvement to concordance when compared to other modifications, and therefore decided not to include dechorionation in the optimized protocol. Henn and Braunbeck [Henn 2011] showed that ZF embryos can be dechorionated at 24 hpf with survival rates of ≥90%; dechorionation of younger stages (<24 hpf) resulted in much lower survival rates. Mandrell et al. [Mandrell 2012] reported the automated dechorionation of 1600 embryos at once at 4 hpf with 2% embryo mortality by 24 hpf and 2% of the embryos malformed at 120 hpf. Automated dechorionation, therefore, seems very promising. However, it is not clear what criteria were used to determine the presence of malformations. It could therefore be that the sensitivity of the method they used was very low. In addition, as these embryos were cultured at 28°C this could also cause an underestimation of the actual effect from dechorionation, as suggested by our data (see Table 8.1). In addition, this machine is not commercially available. Moreover, the 2% mortality rate plus 2% malformation rate caused by dechorionation might be acceptable for high throughput screening, but will not be acceptable for our purpose as it will obscure interpretation of the ZET.

Taking everything into account, it is warranted to leave the chorion intact as dechorionation can lead to increased mortality and abnormality rates. Dechorionation would eliminate the possibility of determining the effects on hatching, is time consuming, and will affect the time at which exposure can start. To assure the embryo is being exposed, we recommend to incorporate bioanalysis of the ZF embryo/larvae into the experimental setup, avoiding false negative and/or positive effects.

8.2.5. Test chamber

The selection of the test chamber was carefully evaluated as some are reported to have deleterious effects in the developing ZF [Busquet 2014, Braunbeck 2005]. The most commonly used test chambers are 96 and 24 wells-plates. The use of 96 wells-plates for ZET experiments represents an attractive option. In some part because it represents an economical approach, since less test volume is needed, and also because more embryos can be fitted in one well plate covering more concentrations in only one well plate.

Ali et al. [Ali 2011] suggested that 96 wells-plates might not be suitable because of the high evaporation rates (9.46% after 96 hrs) which have been reported under static conditions. An almost 10% loss due to evaporation could induce a substantial change in concentration (when only compound or only medium evaporates), which might influence test results. The highest evaporation was observed in outer wells suggesting that avoiding the outermost wells might lessen the evaporation issue, and a semi-static method could additionally lower the losses by evaporation. Also, many

research groups suggest the use of adhesive lids to avoid changes in volume due to evaporation, and to avoid contamination. The use of an adhesive lid in 96 wells-plates might not be suitable as preliminary results have shown high mortality rates of embryos exposed to water in 96 wells-plates covered with adhesive films. Nevertheless, no details were given by the authors [Ali 2011].

Furthermore, it was published by Selderslaghs et al. [Selderslaghs 2009] that 96 wells-plates might induce skeletal malformations. This group observed a higher percentage of ZF with skeletal deformities as the well plate diameter decreased and the exposure period extended. This might be due to the reduction in swimming space leading to malformations. Nevertheless, some research groups use 96 wells-plates for various exposure periods and exposure conditions (static, semi-static), not reporting any deleterious effect of the test chamber in the developing embryo [Padilla 2012]. Therefore, if the amount of test compound is very limited, it could be decided to use 96 wells-plates.

The FET guideline (OECD 236) [OECD 236] recommends the use of 24 wells-plates with 2 mL medium per well as it provides more oxygen supply for the developing ZF. Lammer et al. [Lammer 2009] reported that the oxygen consumption of embryos, especially if they are still inside the chorion, takes place via passive processes depending on the oxygen gradient, and therefore adequate oxygen concentrations in the surrounding medium are essential for appropriate testing conditions. Because of these reasons, we selected 24 wells-plates with 2 mL per well and adhesive lid film, as indicated by OECD 236 [OECD 236], as most optimal test condition. This will solve the risk of big concentration changes due to evaporation, ensures sufficient oxygen supply, and diminishes the occurrence of skeletal deformities by using a bigger volume for the larvae to swim. In addition, examinations of the ZF by microscope are considered more practical using a 24 wells-plate instead of a 96 wells-plate, as the view is much clearer.

8.2.6. Exposure length

The length of exposure is defined by: 1) the period of ZF development that should be covered, 2) the maximum age at which the ZF is still considered non-protected according to the EU Directive on the protection of animals used for scientific purposes [EU Directive 2010/63/EU, 2010], and 3) practicality.

In literature various exposure periods are reported, i.e.: lasting until 48-52 hpf [Busquet 2008, Teixidó 2013], until 72-80 hpf [Fraysse 2006, Hermsen 2011], until 96 hpf [Hallare 2004, van den Bulck 2011] until 120 hpf [Gustafson 2012, Padilla 2012, Ali 2011, Huang 2010, Jomaa 2014] or until 144 hpf [Selderslaghs 2009]. It can be argued that some compounds would need specific embryonic stages to elicit their teratogenic effects, especially if specific receptors/targets need to be present to elude the teratogenic effect. Nevertheless, it appears that most teratogenic compounds can elicit teratogenic effects at very early life stages, as somite formation and heart development seem to be by far the most sensitive structures in the developing ZF to teratogenic compounds [Ducharme 2013]. In the ZF, organogenesis of the main organs is completed by 72 hpf, most organs

are fully developed by 96 hpf, and by 120 hpf the development of the larva is completed [Kimmel 1995, Delvecchio 2011].

Selderslaghs et al. [Selderslaghs 2012] argues that extending the exposure period to 144 hpf was an important parameter for the correct classification of lithium chloride in ZF embryos. However, the embryotoxic effects of lithium on ZF embryos were already noticeable by 72 hpf, but as their classification approach (i.e. teratogenic index as a ratio of LC50/EC50) was unsuitable it resulted in an incorrect classification of this compound at 72 hpf.

In addition, Jomaa et al. [Jomaa 2014] recently published their results on the effect of extending the exposure period for thyroid-active compounds from 72 to 120 hpf. They suggest that extending the exposure period to 120 hpf gives more reliable results for teratogenicity than their initial exposure length of 72 hpf. However, only thyroid-active compound were tested and as production of T4 starts around the time of hatching (72 hpf) [Chang 2012] extending exposure beyond 72 hpf is certainly valuable. However, extending the exposure length to 120 hpf could raise practical issues due to regulation on experimental animals. Therefore, an exposure length between 72 hpf and 120 hpf is suggested to be optimal, maintaining practicality. Because Jomaa et al. [Jomaa 2014] did not evaluate the exposure length of 96 hpf, it is not clear if extending the exposure length to 120 hpf is of added value.

According to the EU Directive on the protection of animals used for scientific purposes [EU Directive 2010/63/EU, 2010], the earliest life-stages of animals are not defined as protected and, therefore, do not fall into the regulatory frameworks dealing with animal experimentation in Europe. Independent feeding is considered as the stage from which free-living larvae are subject to regulations for animal experimentation. Based on extensive analysis and by including criteria such as yolk consumption, feeding, and swimming behaviour, Strähle et al. [Strähle 2012] concluded that ZF larvae at 28.5°C can be regarded as independently feeding from 120 hpf onwards. However, it was reported that the digestive system matured already at 96 hpf, but free swimming and, therefore, the potential to catch prey efficiently, occurred for the majority of ZF after 120 hpf. As this concerned the majority of ZF, it means that some of the ZF will be free swimming already before 120 hpf. As this was determined at 28.5°C, it would be wise not to continue the ZET up to 120 hpf when cultured at 28°C. When using 26°C the ZET could continue until 120 hpf. However, as our proposed study design includes daily evaluations, to exclude morphology assessments in the weekend, we selected an exposure length of 96 hours (i.e. until 100 hpf) as it covers organogenesis, is within the unprotected part of the ZF life, and would allow performance of the ZET during weekdays.

8.2.7. Choice of concentrations

The choice of concentrations being tested is crucial for developmental toxicity assessment in ZF as these should not cause lethality. Standard aquatic toxicology guidelines require that the testing is done up to water solubility or 100 mg/l (whichever is lower). The guideline for the Early Live Stages (ELS) (OECD 210) [OECD 210] sets the maximum to 10 mg/l. This test is a prolonged (i.e. 32 days post fertilisation), sub-lethal toxicity test.

A reasonable approach is to use the Maximum Tolerated Concentration (MTC) and dilutions thereof, as recommended by the guidelines to assess for endocrine disruption (OECD 229 and 231) [OECD 229, OECD 231]. MTC is defined as the concentration causing no more than 10% of acute mortality. It is a concentration at which systemic toxicity does not obscure the more subtle effects (e.g. endocrine, teratogenic).

When data from an acute fish toxicity study is available, then the MTC can be calculated as 1/3 of LC_{50} as recommended by the OECD guidelines. Otherwise MTC can be estimated from a range-finding test performed with concentrations reaching 100 mg/l or up to solubility limit, whichever is lower, with a spacing factor of 10.

The most commonly used maximum spacing factor for the definitive test with fish is 2.2. At least 5 concentrations (in both range-finding and definitive test) should be tested in order to obtain a reliable dose-response curve.

It would be highly recommended to always perform a range-finding test in which bioanalyses and formulation analyses are included. This is crucial for determination of the correct dose levels and most optimal exposure method of the definitive study.

8.2.8. Exposure method

The exposure method is an important factor which could greatly influence the test compound concentration in the medium throughout the experiment. In aquatic toxicology three methods are used: static, semi-static and flow through [OECD GD23, 2000]. The choice of the method is based on the stability of the concentrations during the exposure. A static method represents an economic and fast approach since test solutions are prepared only once at the beginning of the test, using less test compound per experiment. It also optimizes the high throughput of the test by avoiding the time consuming procedure of solution preparation for daily renewals. Static exposure is, however, only allowed in case stability of the concentrations is assured during the exposure. The OECD guideline 236 [OECD 236] suggests that daily refreshments should at least be performed for compounds that show degradation higher than 20% at the end of 96 hours of exposure. Although several research groups use the static method with high predictability rates [Brannen 2010, Gustafson 2012, Selderslaghs 2009], the risk/benefit evaluation suggests a semi-static or flow-through method would represent a better alternative, especially when dealing with volatile or highly hydrophobic compounds in which test concentration is expected to vary during the course of the experiment.

A flow-through method would be ideal when dealing with compounds which have a high demand of oxygen, or highly unstable in aqueous solutions. Lammer et al. [Lammer 2009] developed a flow-through method using 24 wells-plates as test chambers. Although effective, the equipment ended up to be too expensive and laborious.

A prior knowledge about the properties of the compound (e.g. hydrophobicity, volatility) and the stability of the concentration is, therefore, very important in choosing the correct method. If there is no information about stability it is advisable to perform a preliminary stability test in the wells plates. In case it cannot be done, e.g. due to difficulties with measurements, the semi-static exposure is the safest option to proceed with.

Important to know is that on each solution renewal (when using semi-static conditions), the exposure media should be removed with a plastic Pasteur pipette leaving a small amount of solution, to avoid eggs from contact with air and the risk of drying.

8.2.9. Use of solvent

The guidelines for aquatic (eco)toxicology assessments require generation of test solutions, which contain completely dissolved compound concentrations or stable, and homogeneous mixtures or dispersions. The testing of concentrations that disturb the test system should be prevented or avoided, e.g. film of the compound on the water surface or extensive precipitation, flocculation, aggregation or deposition of the undissolved fraction of the compound. If this is not attainable by simple shaking or stirring procedure, the method of preparation can be based on the principles laid down in the OECD Series on Assessment no. 23 (OECD 2000) [OECD GD23, 2000].
Use of solvent is considered by ecotoxicological guidelines as a last option and in the case of endocrine testing and screening is considered highly undesirable (OECD 229, 230, 234) [OECD 229, OECD 230, OECD 234]. However, the advantages of this method of application cause that it is commonly employed in testing with ZF. Here, solvents are used because they facilitate the application of a compound without going through laborious procedures described in the OECD series 23 [OECD GD23, 2000]. Green and Wheeler give an overview of solvent carriers commonly used in aquatic toxicology with all their pros and cons [Green 2013].

Although solvent carriers are commonly used, this should be done after considering all other options. A new, promising, method of application of hydrophobic substances emerging in last years is the use of a passive dosing system. In this method biologically inert o-rings soaked in the compound are introduced individually to test wells and continuously release compound to the solution based on the partition coefficient [Smith 2010].

In the ZET DMSO is, by far, the most common solvent employed. The total concentration of DMSO used by research groups varies in great degree. DMSO has shown to be one of the least embryotoxic solvents for ZF. The embryotoxicity and acute toxicity studies of DMSO done in ZF embryos indicate its safety up to 1% v/v for morphological endpoints [Gustafson 2012, Padilla 2012, Selderslaghs 2009, Hallare 2006, Maes 2012]. However, Weigt et al. [Weigt 2011] reported a doubling of affected embryos in the control groups with 0.5% DMSO compared to controls without (i.e. 4.6% vs 2.0%), in a number of 20 embryos per group. Hallare et al. [Hallare 2006] discovered DMSO to be the most potent inducer of stress proteins when compared to acetone and ethanol in the ZF embryo/larvae, as induction of heat shock protein 70 was noted as low as 0.01% v/v. Significant

reduction of metabolic gene expression is already observed in concentrations as low as 0.1% v/v [David 2012]. It remains unknown whether such effects at the gene expression level represent an effect at the protein level and therefore, an impact in the metabolisms of xenobiotics by the CYP1A family. In addition, Kais et al. [Kais 2013] showed that DMSO affected the distribution of fluorescein in different compartments of the ZF egg (chorion/embryo); DMSO concentrations of 0.1% and 1% decreased the barrier function of the chorion, whereas a concentration of 0.01% DMSO remained ineffective. The effects noted on hsp70 and distribution of fluorescein at 0.01% DMSO was not considered to have an impact on the parameters evaluated in the ZET assay.

Considering all these findings, we conclude that DMSO represents a safe option as a solvent for the ZET when used at a maximum of 0.1% v/v. On a practical level, it should be noted that a maximum of 0.1% v/v DMSO will reduce the number of compounds that can be tested due to solubility limits. DMSO at 0.5% v/v could therefore in these cases only be used when doubling the number of embryos in the experiment. However, still, the most optimal condition would be not to use any solvent.

8.3. MORPHOLOGY SCORING SYSTEM AND ASSESSMENT OF TERATOGENICITY IN THE ZET

Multiple teratogenicity scoring systems have been developed with good predictability rates [Brannen 2010, Padilla 2012, Hermsen 2011, Selderslaghs 2012, Busquet 2008, Weigt 2011, van den Bulck 2011, Teixidó 2013]. For most of these papers, only limited information is given on the exact method used. Only Panzica-Kelly et al. [Panzica-Kelly 2010] provides a very detailed explanation of the used morphological scoring system. However, this system is developed for dechorionated embryos (which they later revisited), is very time-consuming, and only examines once (at day 5 post fertilization) instead of daily. Therefore this method was not considered suitable for our purpose. The method proposed by Hermsen et al. [Hermsen 2011] was more suitable as it uses embryos with intact chorions, provides possibility to evaluate development on a daily basis (providing crucial information for correct classification), gives details on both general developmental delay and teratogenicity, is less time consuming and yet provides complete information. Moreover, its semi-quantitative assessment of specific developmental endpoints supports standardization of the evaluation.

We extended Hermsen's general morphology score (GMS) system [Hermsen 2011] to include assessments for general development on a daily base (instead of only on the last day), and the schedule was extended with one additional day to include exposure for 96 hours (instead of 72 hours). For this additional day, two developmental milestones of ZF larvae at 96 hpf as described by Kimmel et al. [Kimmel 1995] were included in the extended GMS, namely an almost empty yolk sac extension and more apparent protruding mouth.

The advantage for assessments on a daily basis is that it provides crucial information on the development of certain findings over time and might elucidate dose response relationships. Moreover, any morphological abnormalities early in development will be scored, which would be missed when only evaluating on the last day as at that time the malformed embryo would be coagulated, and it will then be scored as mortality instead of malformation. It is of utmost importance to make this distinction. It should especially be realized that coagulation in zebrafish embryos/larvae is a much quicker and more detrimental process than resorption/autolyses in rat and rabbit fetuses.
As already mentioned, this proposed extended morphology scoring system includes two parts: 1) general development, and 2) teratogenicity. Both are described in detail in the following paragraphs (8.3.1 and 8.3.2, respectively).

8.3.1. Description of developmental endpoints of the extended GMS system

Developmental assessment of the embryos is performed at 24, 48, 72, and 96 hours of exposure, using a stereomicroscope. As treatment is started at 4 hpf, these timepoints correspond to 28, 52, 76, and 100 hpf. GMS points are received for each developmental mark, relative to its stage of development. All deviations, for instance incomplete detachment of the tail or lack of pigmentation, will result in a lower GMS. A lower score will correspond to a certain degree of developmental retardation. See Figure 8.2.

Detachment of the tail

Detachment of the tail is scored as 1 if the tail has detached until the yolk extension, which is the part of the yolk-sac that is attached to the tail. A scoring of 2 is given when also the yolk extension is detached. 3 Points for detachment of the tail are given if a full detachment is reached, which is evaluated when the part of the tail distal to the yolk extension is larger than the yolk extension itself. Figure 8.3 gives an overview of examples on how to score for detachment of the tail.

Somite formation

Formation of somites in the embryo/larva is scored either with 0 or 1. A score of 1 is given when somites are visible, either distinct or indistinct (Figure 8.4). However, the latter will be scored as teratogenic, as will be described in the paragraph on deformed body shape (see “8.3.2. Description of teratogenic endpoints”).

Eye development

When no eyes are visible, this developmental point is scored with 0. When eyes are visible, but flat, a score of 1 point is given (Figure 8.5.A). After the appearance of a sphere, the eye development is scored with 2 points (Figure 8.5.B to 8.5.D).

Additionally, the presence of pigment in the embryo’s eyes is scored with +1. So flat eyes (score 1) with pigmentation (+1) receive a total number of 2 points. The maximal scoring for this developmental point is 3.

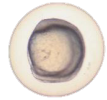
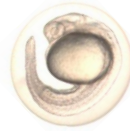
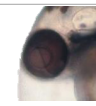
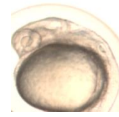
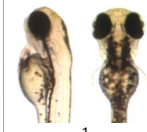
Hours of exposure	Reference in case of retardation	24	48	72	96
Detachment of the tail	 0 1	 2	3	3	3
Somite formation	No = o	Yes = 1	Yes = 1	Yes = 1	Yes = 1
Eye development	1	2	 2 + 1 for pigment	2 + 1 for pigment	2 + 1 for pigment
Movement	No = 0	Yes = 1	Yes = 1	Yes = 2	Yes = 1
Circulation	No = 0	Yes = 1	Yes = 1	Yes = 2	Yes = 1
Heartbeat	No = 0	No = 0	Yes = 1	Yes = 2	Yes = 1
Pigmentation of the head and the body	0	 0	1	 1	1
Pigmentation of the tail	0	0	1	1	1
Yolk extension nearly empty					Yes = 1 No = 0
Pectoral fin	0	0	0	1	1
Protruding mouth	0	0	0	1	2
Hatching	No = 0	No = 0	No = 0	Yes = 1	Yes =1
GMS		**7**	**12**	**15**	**17**

Figure 8.2. Extended general morphology score (GMS) system.

In addition to Figure 8.2, a detailed description on the scoring of developmental endpoints of the GMS system is given in the following paragraphs.

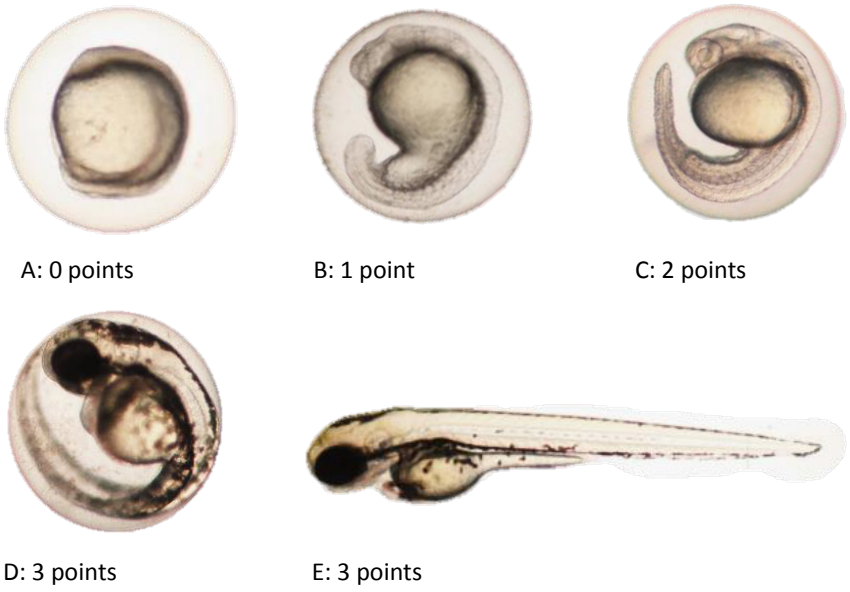


Figure 8.3. Detachment of the tail. A: 0 points. B: 1 point. C: 2 points. D and E: 3 points.

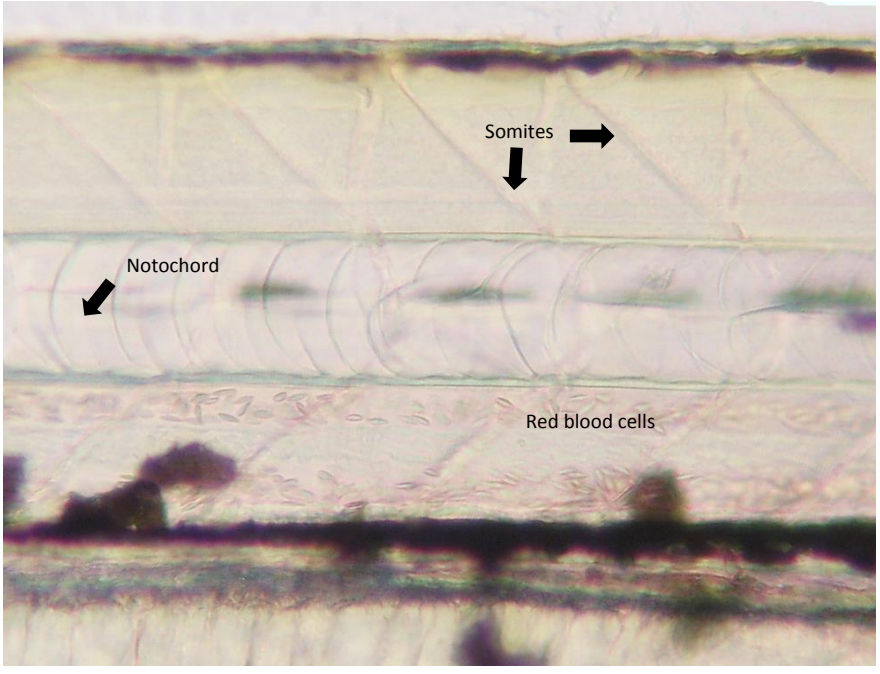
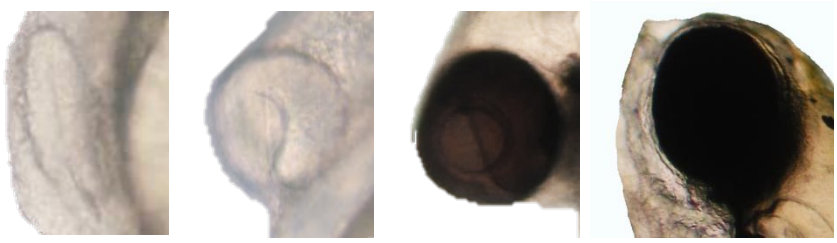


Figure 8.4. Somites.



A: 1 point B: 2 points C: 2 points + 1 pigment D: 2 points + 1 pigment

Figure 8.5. Eye development.

Movement

Movement of the embryo/larva will be scored when the embryo/larva moves its tail or whole body. When pectoral fins have developed, also movement of the pectoral fins is considered to be movement. This endpoint can be evaluated either after spontaneous movement by the embryo/larva, or after stimulation using a plastic pasteur-pipette. In the latter case the embryo/larva is not touched with the pipette, but stimulated by blowing bubbles or establishing a flow in the well. Movement can be scored as absent with 0 points, or present with 1 point.

Circulation

Circulation of blood is scored with 0 points when absent, and with 1 point when present. A blood circulation existing of a few cells is already scored with 1 point. Logically, this developmental point will only be present when also a heartbeat is observed.

Heartbeat

Heartbeat is scored with 0 points when absent, and with 1 point when present. When no heartbeat is visible, a higher magnification (of about 100x) should be used to zoom in on the heart region to look for the presence of a very weak heartbeat. This is done for at least 1 minute until one or more distinct heartbeats are noted.

Pigmentation of the head and body

Pigmentation of the head and body is scored with either 1 point for presence or 0 points when absent. During normal development, this developmental point appears from 48 hpf. Pigmentation is evaluated as present when pigment is visible between the head to the end of the yolk extension, as seen in Figure 8.6.B to 8.6.F. Figure 8.6 depicts pigmentation of the ZF larva with different intensities, and indicates the differentiation between head and body, and tail. Moreover, pigmentation of the head and body is often earlier visible than pigmentation of the tail.

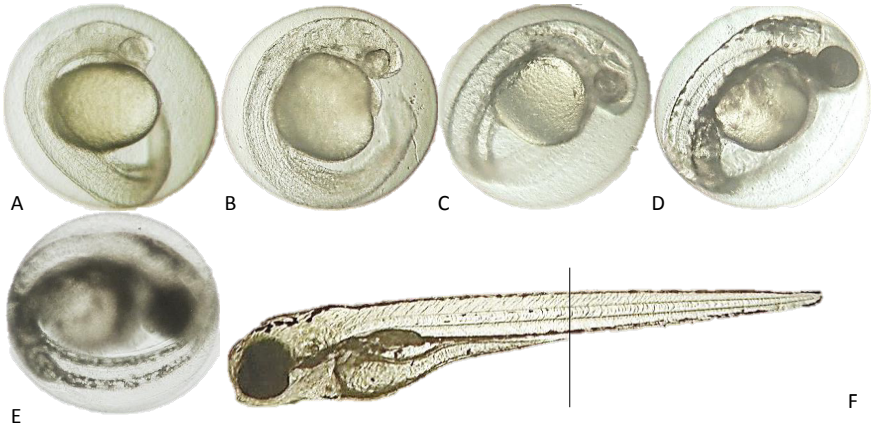


Figure 8.6. Pigmentation of head and body, and tail. A: 0 points. B, C, D: 1 point for head and body. E: 1 point for pigmentation of the tail. F: differentiation between body and tail.

Pigmentation of the tail

Pigmentation of the tail is scored with either 1 point when present, or 0 points when absent. During normal development, this developmental point appears from 48 hpf and is scored similar to pigmentation of head and body. Pigmentation is evaluated as present, with 1 scoring point, when a distinct pattern of pigment is visible from the end of the yolk extension to the tail, as seen in Figure 8.6.F. Different levels of pigmentation in the tail of the ZF larva are illustrated by Figure 8.6.

Yolk extension nearly empty

A nearly empty yolk extension is evaluated as present, with 1 developmental point, or not applicable, with 0 points. This developmental point is only scored at 96 hours of exposure. A nearly empty yolk extension is reached when the thickness of the yolk sac does not exceed the heart and a space is visible where the swimming bladder will be situated. Figure 8.7 shows examples of a larva with a nearly empty yolk extension (A) versus one whose yolk sac has not yet nearly emptied.

Pectoral fin

The visibility of pectoral fins is scored with 0 points when not present, and with 1 developmental point when present. Pectoral fins originate approximately in between the heart and the liver extending posteriorly along each side of the larva. The presence of pectoral fins is also scored with 1 point when only one of the two pectoral fins is visible. This developmental point normally appears from 72 hpf.

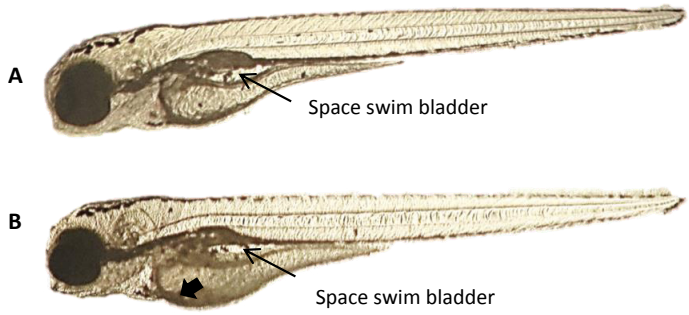


Figure 8.7. Empty yolk extension in larva at 96 hours of exposure. A: 1 point for nearly empty yolk extension. B: 0 points for nearly empty yolk extension. Yolk sac B exceeds the heart (see thick arrow). Both show spaces where the swimming bladder will be situated.

Protruding mouth

A protruding mouth is absent (score is 0) when no lower jaw is seen or when the lower jaw only reaches until about half of the eye (at the level of the pupil). The presence of a protruding mouth is scored with 1 developmental point when the lower jaw of the embryo/larva reaches throughout the complete eye. A score of 2 points for a protruding mouth is given when the lower jaw reaches to the front of the eye or further. Figure 8.8 gives a scheme on when to score for this developmental point.

Figure 8.8. Protruding mouth in ZF larva: 0 points when the lower jaw is not present or when the lower jaw reaches only till half of the eye. 1 Point is given when the lower jaw reaches from half of the eye to the front of the eye. 2 Points are given when the lower jaw reaches at least until the front of the eye.

Hatching

Hatching is scored with 0 points when not applicable, and with 1 point when applicable. This developmental point normally occurs after 72 hpf. When already present at 48 hpf, it will not be scored but should be noted for aid in further evaluation. The larva is hatched when the body is not surrounded by chorion. The age of 72 hpf, rather than hatching, determines when the embryo becomes a larva [Westerfield 2000]. A gap in the chorion, with the chorion still surrounding the body of the embryo/larva is not evaluated as hatched. However, the latter observation should be recorded, as the gap can influence the exposure of the embryo or larva to the test compound and surrounding medium.

8. 3.2. Description of teratogenic endpoints

The second challenge of the scoring system was to select the correct teratogenic endpoints to assess for. This is a crucial part, since these endpoints will have a direct impact on the throughput of the test, but more importantly, these are directly linked to the predictability of the test [Panzica-Kelly 2010].

For this part of the setup, an evaluation of several teratogenic endpoints assessed in reported teratogenic scoring systems was done, with special emphasis on those endpoints which showed to have good prediction values when classifying teratogenic compounds. Prominence was given to skeletal deformities and heart and yolk edema, as meta-analysis studies have found these endpoints to be the most common toxic targets of teratogenic compounds and good predictors of teratogenicity [Ducharme 2013].

As a basis, Hermsen's teratogenic assessment system [Hermsen 2011] was taken into account and some changes and adjustments were made based on in-house experience and literature. For example, as part of malformation of the head (#1 in Table 8.2), jaw structure assessment and eye integrity were included as these endpoints are proven to be common targets of teratogenic compounds facilitating compound classification [Brannen 2010, Panzica-Kelly 2010]. Furthermore, malformation of the heart (#4 in Table 8.2) is usually caused by edema, so both pericardial edema and heart malformation accounted as one single endpoint and not as separate ones. In addition, deformed body shape (#5 in Table 8.2) is usually caused by chorda abnormalities, such as a modified chorda structure, scoliosis or rachischisis. Therefore, these endpoints were merged and were assessed as one. Furthermore, somite deformations are also assessed as deformations of the body (#5 in Table 8.2), as somite formation is important for assessing compound toxicity. Moreover, yolk sac edema and yolk deformations were also merged as one endpoint (#6 in Table 8.2), as in most cases, yolk sac deformations are caused by yolk edema.

Teratogenic assessment of the embryos/larvae is performed at 96 hours of exposure (i.e. 100 hpf), using a stereomicroscope. The teratogenic endpoints assessed are depicted in Table 8.2 and are scored as present or absent.

Table 8.2. Teratogenic endpoints assessed in the ZF embryo/larva.

	Teratogenic endpoints	Examples
1	Malformation of the head	Missing structures such as the jaw and the eyes. Uneven eye size or abnormal eye shape, abnormal head shape, edema etc.
2	Malformation of the sacculi/otoliths	Missing structures, duplicated structures or uneven size or abnormal shape.
3	Malformation of the tail	Kinked tail or malformed tail fins.
4	Malformation of the heart	Irregular shape due to edema or aplasia. Abnormal heartbeat.
5	Deformed body shape	Scoliosis, rachischisis, absence, indistinct or malformed morphology of the notochord or (one or more) somites.
6	Yolk deformation	Edema or malformations.

In addition to Table 8.2, a detailed description on these parameters is given below.

Malformation of the head

A malformation of the head is established when structures are missing, such as the jaw, and eyes. Also an uneven or small eye size, abnormal shape of eyes, abnormal head shape, etc. are scored as positive for malformation of the head. An abnormal head shape could for instance be a very flat head, when the forehead is in line with the frontal edge of the eyes, or a bulb on the forehead or back of the head.

Malformation of the sacculi/otoliths

A larva is positive for malformation of the sacculi/otoliths when one or both of these structures are missing, duplicated, or have an uneven size or abnormal shape. Also enlarged sacculi and/or otoliths, or not well-defined structures are scored as teratogenic.

Malformation of the tail

The tail is scored as malformed if a kink is seen in the tail (Figure 8.9). This is mostly seen at the tip of the tail and can be either up-lifted or down. Also malformed tail-fins are scored as a malformation of the tail.

Figure 8.9. Kink in tail of larva at 96 hpf.

Malformation of the heart

Teratogenic endpoints of the heart are an irregular shape due to edema or aplasia, and an abnormal heartbeat; to slow, to fast, or irregular compared to the negative control.

Deformed body shape

Malformations of the body are abnormalities in the neural tube, absence, indistinct or malformed morphology of the notochord or (one or more) somites. Figure 8.10 illustrates some examples of larvae with deformed body shapes. Additional malformations are disclosure of the back (rachischisis) or malformed skeleton (scoliosis).

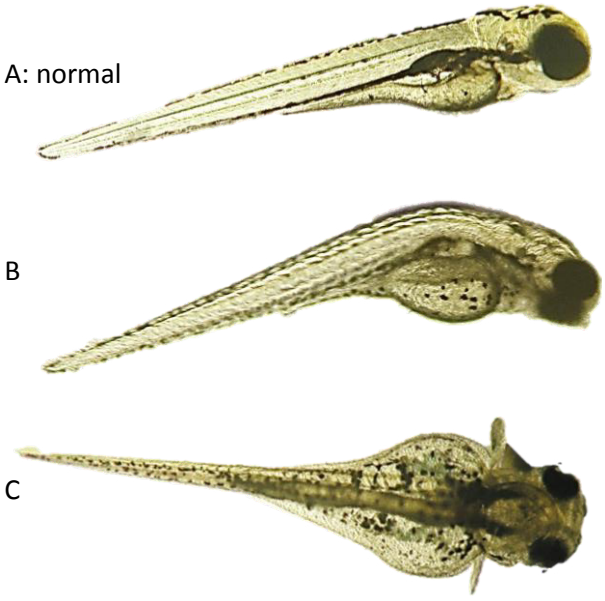


Figure 8.10 Larvae at 96 hours of exposure with deformed body shapes (B and C) versus normal larva (A). Larva B hatched at least 24 hours before this picture was taken and still showed bending of the body, and is therefore evaluated as positive for deformed body shape. Larva C has a deformed head with small eyes, and a deformed body shape with enlarged yolk sac and large malformed pectoral fins.

Bending of the body must be evaluated in addition to the information on hatching time. If no hatching was seen at 72 hours of exposure, bending of the body will not be considered a teratogenic finding. Bending of the body after hatching is illustrated by Figure 8.11.

Yolk deformation

The yolk sac is scored as teratogenic when edema is established, as illustrated by Figure 8.12. Also bulbs or other malformations that interfere with the normal rounded shape of the yolk are scored as teratogenic endpoints.

Figure 8.11. Deformed body shape of larva during process of hatching at 96 hours of exposure.

Figure 8.12. Severe edema of yolk sac at 96 hrs.

8.4. CLASSIFICATION OF TERATOGENICITY

The teratogenic index (TI) is a popular approach to classify compounds according to their teratogenic potential [Brannen 2010, Selderslaghs 2012, Weigt 2011]. The TI value can be obtained by different calculations (e.g. LC_{50}/EC_{50}, LC_{25}/NOAEL) and is greatly influenced by the pre-set value that needs to be reached for being classified as a teratogenic compound (e.g. $LC_{50}/EC_{50} \geq 2$, LC_{25}/NOAEL ≥ 10). Evaluating the papers using these approaches, the LC_{50}/EC_{50} ratio is the least appropriate as, even though embryotoxic effects are observed at earlier time points, if either the LC_{50} or the EC_{50}

are not reached within the exposure period, the compounds are automatically classified as non-teratogens [Selderslaghs 2009].

Another, and more promising, method is the benchmark dose (BMD) approach [Slob 2002, Piersma 2008], in which the benchmark concentration (BMC) at a predefined benchmark response (BMR) will be calculated using a fitted dose-response curve. It is a very practical approach as it allows obtaining a concentration at which a relevant change in normal morphology can be expected. It is important to consider the selected percentage of decrease in normal morphology (e.g. 5 or 10%) when calculating the BMC for teratogenicity (BMC_t). The scientific committee formed by EFSA concluded in their guidance document that the BMD makes extended use of available dose-response data and provides a quantification of uncertainties in the dose-response data [EFSA Guidance 2009]. This is not done by the TI approach.

Using the proposed morphology scoring system, we would determine a BMD for developmental delay at 24, 48, 72 and 96 hours and a BMD for teratogenicity (one for number of embryos with malformations, and one for total number of malformations with a maximum of six teratogenic endpoints per embryo).

Currently, for both approaches (TI and BMD) there are two shortcomings which are important to overcome in future research. The first is the pre-set value (e.g. TI ≥2 or $BMC_t \geq 5\%$) at which teratogenicity is defined. More research is needed to define these cut-off values. The second drawback is the omission of true compound uptake. It would be of great value to evaluate the TI or BMD together with its body burden by adding bioanalysis.

For example, Gustafson et al. [Gustafson 2012] showed fluconazole (a definitive human teratogen) to be non-teratogenic in ZF embryos; however bioanalysis showed a very low uptake (5.1% uptake as % of nominal well solution). On the other hand, loratadine (not teratogenic in humans or rodents) showed to be teratogenic in ZF embryos; bioanalysis showed a very high uptake (4500%). Therefore, the body burden should be taken into account when evaluating presence or absence of teratogenic effects. The main goal would be to define the minimum and maximum threshold values. Van den Bulck et al. [van den Bulck 2011] used a pre-determined 50 ng/larvae as maximum body burden, which was considered successful by Doshna et al. [Doshna 2009] but showed to be too high for the tested compounds by Van den Bulck et al. [van den Bulck 2011]. In addition, Ball et al. [Ball 2014] implemented the following strategy: higher dosing concentrations were employed with compounds which demonstrated no lethality, morphological malformations and low uptake (< 5%). So, they defined the minimum level of body burden as 5%.

Overall it is concluded that the best classification approach would be the BMD in combination with body burden analyses. The BMD provides a concentration at which a relevant change in normal morphology can be expected, which allows comparison of teratogenicity severity between compounds. And body burden analyses is needed to define false positive and false negative results.

8.5. DISCUSSION AND CONCLUSION

In summary, based on the extensive literature review and in-house experience in combination with the OECD guideline for the FET, the following most optimal ZET conditions are proposed:

	Most optimal condition	Considerations
Selection and number of eggs	Only select healthy looking eggs showing no signs of coagulation. Use at least 30 embryos per concentration.	• Signs of coagulation at 4 hpf are: presence of bubbles around the animal pole and asymmetry of the embryo. • Numbers are based on the OECD validation of the FET test.
Temperature	26 °C	• Pilot results indicate 28°C could accelerate development, thereby missing teratogenic effects as the ZF are coagulated at the time of assessments or no effects are induced as the window of exposure was too limited. • It will provide a worst-case scenario. • More research on this is needed.
Chorion	Intact	• In general, the chorion does not affect the predictability; bioanalyses should be incorporated to assure ZF exposure. • Dechorionation will delay start of exposure and could lead to malformations.
Test chamber	24 wells-plate (2 mL per well) in combination with self-adhesive film cover	• Possible high evaporation rate in 96 wells-plate could induce variability in dose concentration. • Limited well space/volume in 96 wells-plate could induce skeletal malformations. • To avoid contamination (one concentration per plate is used with 24 well plates).
Exposure length	Start at 4 hpf	• This is the end of the blastula period, corresponding to implantation in rats and rabbits (start of maternal exposure in developmental toxicity studies).
	End at 100 hpf (96 hours of exposure)	• 96 hours is sufficient to classify a compound when using the correct classification approach. • Covers entire organogenesis • Complies with the EU Directive on the protection of animals [EU Directive 2010/63/EU, 2010].
Choice of concentrations	Based on range-finding test (including bioanalyses and formulation analyses)	• When data from an acute fish toxicity study is available the MTC can be calculated as 1/3 of LC_{50}. • Otherwise MTC can be estimated from a range-finding test performed with concentrations reaching 100 mg/l or up to solubility limit, whichever is lower, with a spacing factor of 10. • At least 5 concentrations (in both: range-finding and definitive test) should be tested in order to obtain a reliable dose-response curve (maximum spacing factor for the definitive test ≤2.2).

Continued:

	Most optimal condition	Considerations
Exposure method	Options:	• The exposure method should guarantee stable test concentrations throughout the experiment:
	- Static	• Economic and fast option when compound is known to be stable during the exposure period.
	- Semi-static	• More time-consuming option, but pivotal when compound degradation >20%, the compound has high demand of oxygen, or when compound stability is unknown.
	- Flow though	• Economic and fast alternative for the semi-static method.
Use of solvent	0.1% v/v DMSO (only if needed)	• Most common solvent in ZET with established safe use up to 0.1% v/v. • 0.5% v/v DMSO could be used (with less soluble compounds) only when using twice the number of embryos.
Morphology scoring system	Extended general morphology score with daily assessments	• It is robust, efficient and includes daily assessments of growth and end-phase teratogenicity. • Daily assessment are crucial for distinction of malformations versus mortality.
Classification of teratogenicity	BMC with body burden	• BMC is calculated using a fitted dose-response curve to obtain an effect concentration allowing comparison of teratogenicity severity between compounds. • Using a Teratogenic Index, if either the LC_{50} or the EC_{50} are not reached within the exposure length, the compounds are automatically classified as non- teratogenic; even when embryotoxic effects are noted. • Bioanalyses should be added for determination of false positives or false negatives.

From ZF literature, a wealth of information is available to select the test conditions that are considered to impact the outcome of the ZET. These data also showed that methodologies of the ZET used by the different groups vary greatly, and harmonization is needed to make the ZET applicable as successful alternative in future safety assessment.

Before harmonization can take place, first the different test conditions need to be further characterized and validated. Some conditions were already described in detail, and based on literature and/or guideline experience, the most optimal test condition could be proposed with confidence (e.g. choice of concentrations, exposure method, and use of solvent). But for other parameters, very limited information was available to definitely propose the most optimal ZET condition (e.g. temperature, and threshold value for teratogenicity). Although the FET has another endpoint (mortality instead of teratogenicity), its study design has been thoroughly validated and can be of imperative guidance to determine other ZET test conditions (e.g. test chamber, and number of eggs). To achieve all characterizations and validations, it would be essential to perform the ZET according to the proposed study design with several positive and negative control compounds, and

then change one of the test conditions (e.g. perform concurrent experiments at 26°C and 28°C in which temperature is the only modification). This should be done for all test conditions.

Based on the presented data, it is strongly recommended to always include bioanalysis to asses ZF exposure. This would aid in distinction of false negative or false positive results. Further investigation is needed to define at which internal concentration false results (positive or negative) can be expected.

In conclusion, this paper provides a structured overview on the pro's and con's of the different test conditions, and can therefore be used as a basis for further development and optimization of the ZET. The proposed most optimal ZET conditions are based on extensive literature evaluation and partial in-house experience. However, it should be emphasized that first characterization of the test conditions is needed, followed by experimental validation by different groups on a wide variety of compound classes to confirm its optimal prediction, before harmonization can be achieved. Taken together, the ZET is an exceptionally promising alternative for replacement of or addition to mammalian testing as now also the ICH is considering the introduction of the ZET in standard developmental toxicity guidelines [ICH S5(R3) Draft January 2014] as an integrated testing strategy. To achieve implementation for safety assessment, global harmonization and improvement of the protocol is essential, and we therefore encourage all users to discuss and publish their methods and results to obtain this goal.

Acknowledgements
The authors would like to thank Jeffrey Charlap for his valuable and dedicated contributions.

Chapter 9

Visualizing compound distribution during zebrafish embryo development: the effects of lipophilicity and DMSO

Coco de Koning, Manon Beekhuijzen, Marysia Tobor-Kapłon, Selinda Buitenweg, Dick Schoutsen, Nico Leeijen, Beppy van de Waart, Harry Emmen.

WIL Research Europe B.V., Hambakenwetering 7, 5231 DD, 's-Hertogenbosch, The Netherlands.

Published in: Birth Defects Research (Part B). 2015 Dec;104(6):253-72. doi: 10.1002/bdrb.21166. Epub 2015 Dec 11.

ABSTRACT

The predictability of the zebrafish embryo model is highly influenced by internal exposure of the embryo/larva. As compound uptake is likely to be influenced by factors such as lipophilicity, solvent use, and chorion presence, this article focuses on investigating their effects on compound distribution within the zebrafish embryo.

To visualize compound uptake and distribution, zebrafish embryos were exposed for 96 hours, starting at 4 hours post fertilization, to water-soluble dyes: Schiff's reagent (logP -4.63), Giemsa stain (logP -0.77), Van Gierson stain (logP 1.64), Cresyl fast violet (logP 3.5), Eosine Y (logP 4.8), Sudan III (logP 7.5), and Oil red O (logP 9.81), with and without 1% dimethyl-sulfoxide (DMSO). Three additional compounds were used to analytically determine the uptake and distribution: Acyclovir (logP -1.56), Zidovudine (logP 0.05), and Metoprolol Tartrate Salt (logP 1.8). Examinations were performed every 24 hours.

Both methods (visualisation and specific analysis) showed that exposure to higher logP values results in higher compound uptake. Specific analysis showed that for lipophilic compounds >90% of compound is taken up by the embryo. For hydrophilic compounds, >90% of compound within the complete egg could not be associated to embryo or chorion and is probably distributed into the perivitelline space. Overall, internal exposure analyses on at least two occasions (i.e. before and after hatching) is crucial for interpretation of zebrafish embryotoxicity data, especially for compounds with extreme logP values. DMSO did not affect exposure when examined with the visualisation method, however, this method might be not sensitive enough to draw hard conclusions.

Keywords
zebrafish embryos (danio rerio), compound uptake, internal exposure, lipophilicity, logP, solvent, dimethyl-sulfoxide (DMSO), 3R's

9.1. INTRODUCTION

9.1.1. The predictability of tests using zebrafish embryos is highly influenced by compound uptake

Early life stages of (up to 120 hours post fertilisation) Zebrafish (*Danio rerio,* further referred to as ZF) are considered a new promising alternative animal model for developmental toxicity studies. Recently, the ZF embryotoxicity test (ZET) was proposed as update in the ICH S5 guideline for preclinical testing of pharmaceuticals. For some time now, discussions are ongoing on the added value of a second species in developmental toxicity assessment (i.e. the use of both rats and rabbits in these preclinical tests). Future perspectives are aiming for a shift from the current two-species testing paradigm to an integrated testing strategy (ITS) consisting of a combination of one or more *in vitro* or non-mammalian *in vivo* embryotoxicity tests and an *in vivo* embryo-foetal development study in one species. The draft concept paper mentions that "a combination of in vitro assays, in silico prediction, and ZF assays is advocated rather than relying on the use of one stand-alone method".

The zebrafish (ZF) model is not regulated by current legislations on animal welfare up to 120 hours post fertilization (hpf). This is because a licence is required on fish from the time at which they become capable of independent feeding, which for the majority of ZF is at 5 days post fertilization (dpf) [Fleming, 2007]. Therefore, the use of ZF embryos is considered a replacement of current animal experiments for toxicity and teratogenicity.

Moreover, where present models for teratogenicity, usually rats and rabbits, require sacrifice of the dam to obtain the embryos for evaluation of developmental toxicity at a single time point, the ZF model allows continuous developmental assessment without sacrifice of the maternal fish. ZF organogenesis is completed within 72 hpf and its development is similar to mammalian embryos [Westerfield, 1995]. Moreover, ZF embryos remain transparent throughout organogenesis into the larva stage. This facilitates detailed evaluation of structures and organ systems in an intact non-mammalian vertebrate embryo at all time points during its complete embryogenesis. Not only the rapid organogenesis of the transparent ZF embryo, but also the simplicity of maintaining and breeding the adult ZF, makes this animal model ideal for developmental toxicity screening.
The predictability of the ZF model, however, is highly influenced by compound uptake and thus internal exposure. Therefore, the mentioned advantages are counterbalanced by uncertainties about internal exposure and related sensitivity of the system. It is thought that the chorion is an important barrier for penetration of compounds until hatching. This means that several physico-chemical properties of compounds tested with ZF embryos are important for generation of valid results. Important factors are size, molecular weight, lipophilicity, and the presence or absence of active transport into and within the embryo.

9.1.2. LogP and chorion presence are factors that could influence compound uptake within the embryo

The chorion of ZF embryos contains pores that have an average size of 0.17 µm2 [Rawson, 2000; Cheng 2007; Lee 2007], which restricts uptake of compounds exceeding the size of 3 kDa or 3000 g/mole [Creton, 2004]. Moreover, the uptake of compound through the chorion and into the embryo is thought to be enhanced by use of solvent within the test solution. Several studies found an effect of dimethyl sulfoxide (DMSO) on compound uptake and speculated that it might affect the predictability of the test [Hallare, 2006; David, 2012; Kais, 2013].

One of the most important factors for internal exposure is thought to be lipophilicity expressed as logP value, of a compound. The logP has already been linked to bioavailability of chemicals in adult fish [Könemann, 1980; Gestel, 1985; Connell, 1988; Tran, 2007; Fu, 2009; Schreiber, 2009; Arnot, 2010; Peterson, 2011]. In addition, chemicals with higher logP, and thus more lipophilic, tended to be more often toxic to the developing ZF than chemicals with lower logP [Padilla, 2012]. Only few chemicals with logP below 0 were found to be toxic in ZF embryos [Steinberg, 2013]. Therefore, it would be interesting to investigate the influence of logP on internal exposure in the ZET as especially during the first days of ZF development the chorion is present and the composition of the embryo changes over time.

In this study, we investigated the effect of lipophilicity on internal exposure and distribution of a compound within a ZF embryo and larva. In order to investigate and visualise the effect of lipophilicity on the uptake and distribution of compounds, we exposed ZF embryos to compounds with a wide range of logP values ranging from hydrophilic (LogP: -4.63) to lipophilic (LogP: 9.8). In order to exclude a possible effect of molecular weight on compound uptake, molecular weights of all compounds tested were well below the chorion pore size limit of 3000 g/mole. Seven of these were dyes that allowed visualisation of the compound's distribution within the embryo or larva. In addition, three compounds with different logP values were used, of which the concentrations within the complete egg, embryo, chorion, and larva were measured over time using Liquid Chromatography- Mass Spectrometry (LC-MS). The latter allowed determination of distribution ratios within the ZF egg. Additionally, we investigated the effect of DMSO on compound uptake and distribution within ZF embryos using dyes.

9.2. MATERIALS & METHODS

9.2.1. Definitions

Egg: fertilised egg including embryo, perivitelline space (PVS) and chorion Embryo: organism before hatching Larva: organism after hatching.

9.2.2. Experimental set-up

ZF eggs were acquired from Zodiac; Wageningen UR, The Netherlands. Compound uptake and distribution was determined within ZF during 96 hours of exposure using two methods: (1) visualisation after exposure to seven dyes with a wide range of logP values (-4.63 to 9.81), and (2) measurements of internal concentrations with LC-MS bioanalysis after exposure to three compounds with different logP values (-1.56 to 1.8). For all experiments we applied procedures based on suggestions for harmonization of the ZET, as previously described by our group [Beekhuijzen, 2015]:

- Test chambers were pre-soaked in test solutions for at least 24 hours to avoid test compound concentration losses due to compound sorption to the well walls.
- Pre-selection of eggs was performed using an Olympus CK2 ULWCD 0.3 light-microscope. Only fertilized eggs with a round chorion and no signs of coagulation, were selected. Selected eggs were immersed in the test solutions upon reaching the gastrulation stage at 4 hpf.
- ZF embryos were cultured at 26 °C .
- Exposure media were refreshed every day.
- Test solutions were preheated in a waterbath to 26°C before use.
- During each refreshment, the exposure media was removed with a plastic pasteur-pipette leaving a small amount of solution, to avoid eggs from contact with air and the risk of drying.
- pH value, temperature, hardness, oxygen concentration and conductivity within the media, and lux and temperature within culture, were kept constant and within values as described in Table 9.1.
- At any time during the experiment, coagulated embryos were recorded as dead and removed.

9.2.3. Compounds tested

Table 9.2 provides detailed information on compounds used in this study. The nominal concentrations tested in experiments are summarized in Table 9.3; for Acyclovir, Zidovudine, and Metoprolol these were actual concentrations, as verified with LC-MS on media samples.

Acyclovir, Zidovudine, and Metoprolol Tartrate Salt (Metoprolol) were used for compound uptake evaluation by LC-MS bioanalysis. All of these compounds were obtained from Sigma-Aldrich (Zwijndrecht, The Netherlands). Concentrations of each compound were prepared in adjusted ISO-medium (ISO 6341, 1996) (purified water containing $MgSO_4 \cdot 7H_2O$ 88.8 mg/L, $CaCl_2 \cdot 2H_2O$ 211.5 mg/L, $NaHCO_3$ 46.7 mg/L and KCl 4.2 mg/L). The concentrations were based on toxicity studies to determine sub-lethal concentrations [Beekhuijzen, 2015].

The dyes used to visualize compound uptake and distribution were from Klinipath (Klinipath B.V., The Netherlands), only Giemsa staining was from Merck (Merck KCaA, Germany). Next to visualization of compound uptake and distribution within the ZF, we also investigated the effect of 1% DMSO on compound uptake and distribution within the ZF model. The compounds used in this, with and without 1% DMSO, are reported in Table 9.3. ZF embryos exposed to solutions without 1% DMSO were used as controls. The concentrations of these dyes were based on the most op-

timal visibility of colour uptake within the embryo, with minimal toxic effect, from pilot experiments. Solutions were diluted with ISO-medium.

Table 9.1. General test conditions for ZET

Lux intensity	550-1080 lux
pH	6.5-8.5 (difference < 1.5)
Temperature	26 ± 0.5 °C
Oxygen saturation	≥ 80 %
Hardness	100-300 mg/L CaCO3
Test chamber	polystyrene 24 wells plate (Greiner-Bio One)
Lid used	self-adhesive; clear polyester sealing tapes (EASYseal™ Greiner-Bio One)
Volume of exposure	2 ml
Exposure method	Semi-static
Photoperiod	14:10 h light
Age of eggs at start of the experiment	4 hpf
Number of eggs per concentration	10 (for LC-MS) or 20 (for visualization)
Number of internal controls per plate	4
Duration of experiment	96 hrs
Chorion presence	Present
Euthanasia	Immersion in ethylene glycol 1.2 % solution, followed by freezing at -20 °C
Euthanasia for bioanalysis	Snap freezing in liquid nitrogen, followed by freezing at -20 °C

Table 9.2. Details on compounds tested

Compounds for bioanalysis	*logP value*	*mW (g/mole)*	*Water solubility according to MSDS*	*CAS number*
Acyclovir	- 1.56	225.2	1 mg/ml	59277-89-3
Zidovudine	0.05	267.2	0.1-0.5 mg/ml	30516-87-1
Metoprolol Tartrate Salt	1.8	267.36	Very	56392-17-7
Compounds for visualization				
Schiff's reagent	- 4.63	191,91	17.2 mg/ml	507-09-5
Giemsa stain	- 0.77	291,8	Very	51811-82-6
Van Gierson stain	1.64	229,1	Infinitely	88-89-1
Cresyl fast violet	3.5	321,34	1 mg/ml	6625-37-2
Eosine Y, solution A	4.8	647,89	400 mg/ml	95917-83-2
Sudan III	7.5	456,54	0.1 mg/ml	85-86-9
Old red O	9.81	408,49	Slightly	1320-06-5

Table 9.3. Concentrations of compounds tested

Compound	*Tested concentrations (µM)*	*Tested concentrations with 1% DMSO (µM)*
Acyclovir	0.1, 1, 10, 100, and 1000	Not applicable
Zidovudine	0.1, 1, 10, 100, and 1000	
Metoprolol Tartrate Salt	0.1, 1, 10, 100, and 1000	
Schiff's reagent	5000 and 10,000	10,000
Giemsa stain	250 and 500	100
Van Gierson stain	5000 and 10,000	10,000
Cresyl fast violet	5000 and 10,000	5000
Eosine Y, solution A	250 and 500	100
Sudan III	1000 and 2000	1000
Oil red O	1000 and 2000	1000

For preparation of test solutions without DMSO, the appropriate amount of stock was weighed, after which ISO-medium was added to mix using a vortex until a clear solution was obtained, which was evaluated by naked eye. A clear solution is defined as a solution that is transparent and contains no undissolved components. It is possible that the test material was not completely dissolved but formed micelle or suspension not visible to the naked eye. To obtain different concentrations, the highest concentrations was prepared and diluted with ISO-medium. For preparation of test solutions with DMSO, test compound was dissolved in DMSO prior to adding ISO-medium, after which a vortex was used to obtain a clear solution.

9.2.4. Method for visualization of compound uptake using dyes

For the visualization of compound uptake and distribution, pictures were taken every 24 hours. To this end, ZF emrbyos/larvae were transferred from the exposure medium to ISO-medium in order to obtain clear pictures. Pictures of complete eggs, embryos (after removal of chorion) and chorion were taken separately. In the result section, pictures of the representative embryos and larvae are presented. For comparability between pictures, the microscope- and camera-settings were kept constant between photos. A Canon IXUS 240 HS camera was used for pictures, setting for artificial TL-light gave the best pictures. The white balance of the photos (calibrated to background) was adjusted batch-wise with Adobe Lightroom 3.0. This correction removed the effect of the microscope's lamp (yellow colour). It should be noted that the results from visualization are not quantitative but rather qualitative and are only assessed by eye, no other technique was used.

9.2.5. Method for bioanalysis with LC-MS used to measure internal exposure

Bioanalysis was performed to obtain information of compound uptake and distribution within the egg, chorion, embryo, and larva. For determination of the amount of compound taken up by the embryos before hatching, embryos were dechorionated. Chorion samples were also taken. Dechorionation of the embryos was performed in ISO-medium using tweezers, as described by Henn and colleagues [Henn, 2011]. Samples were taken daily, starting at 24 hours of exposure, by

pooling 15 complete eggs, chorions, embryos, or hatched larvae. After washing three times with ISO-medium, the collected samples were sucked dry using a pipette, and frozen in liquid nitrogen to prohibit outflow of test compound and provide instant death to the embryos. All samples were collected into 0.5 ml Eppendorf tubes. These were stored at -20 °C until analysis, which was maximally 2 months of storage for experiments described.

At the time of analysis, samples from complete eggs, embryos, and larvae were thawed and homogenized by maceration, using a metal spatula with a blunt ending. Maceration was not needed for the chorion samples. Solvent precipitation was used to purify samples from proteins. Therefore, 250 µl acetonitrile (ACN) (UPC/MS grade, BioSolve) was added to these samples. The samples were incubated for 30 minutes while being vortexed at room temperature. After centrifugation for 15 minutes at 21885 rcf(xg) (Heraeus Multifuge 3 S-R, ID-Nr. 20057915, Kendro Laboratory Products, Germany) at 4 °C, supernatants containing all of the test compound extracted from the samples were taken and put into 96 wells plates to dry under nitrogen at 50°C for 30 minutes. In addition, the concentration of compound in the medium-control samples was also measured. For this, 50 µl of each medium sample was taken and put into 96 wells plates to dry. Eluents (50 µl) was then added to all samples before analysing by LC-MS using a gradient.

The total amount of test compound per sample was calculated using calibration curves measured at the same time as the samples. The amount of test compound per complete egg, chorion, embryo, or larva was calculated by dividing the mean amount of test compound by the amount of eggs, chorions, embryos, or larvae present in the sample. The obtained value was then adjusted for weight (to ng/mg tissue) in order to correct for size differences in the growing embryo over time. Calculations were performed using Graphpad Prism™ software (La Jolla, CA) with the aid of Microsoft Excel 2010.

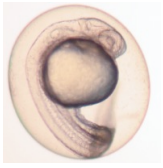
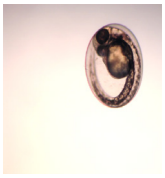
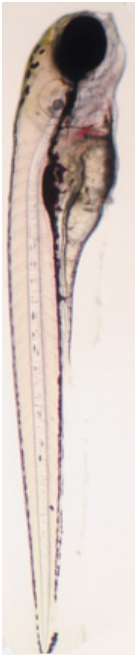
The relative amount of test compound within the complete egg, chorion, and embryo was calculated as percentages of the total amount (complete egg (ng) = chorion (ng) + embryo (ng) + PVS (ng) = 100%).

9.3. RESULTS

9.3.1. Visualization of compound uptake and distribution within the zebrafish embryo related to logP value

Table 9.4. gives the overview of normal colouration of the ZF embryo during development in ISO-medium at 26°C. As shown here, the ZF embryo already has a somewhat yellowish appearance, which increases in the larva stage. Furthermore, tables 9.5 to 9.11 depict the ZF embryos and larvae during 96 hours exposure to the seven dyes.

Table 9.4. Normal colouration of zebrafish embryos and larvae during 96 hours of culture.

	24 hrs (28 hpf)	48 hrs (32 hpf)	72 hrs (76 hpf)	96 hrs (100 hpf)
Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.				Not applicable
				

Eggs exposed to pink Schiff's reagent (logP: -4.63) appear pink and become more pink in time, while the embryos within these eggs are not pink (Table 9.5). The colour of the egg was due to a pink chorion, rather than a coloured PVS as apparent during manual dechorionation. However, after 96 hours of exposure a pink colouration is seen in the yolk sac of the larva. This effect becomes more apparent in the higher concentration, which indicates this effect is concentration-dependent. Moreover, no toxic or teratogenic effects were seen in embryos/larvae exposed to Schiff's reagent.

Table 9.5. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Schiff's reagent.
LogP: -4.63. Colour staining: pink.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation where applicable.

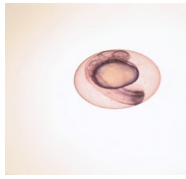
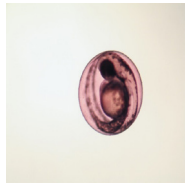
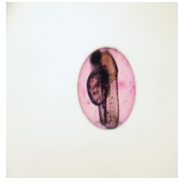
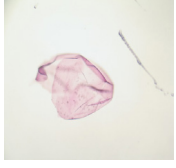
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
5000 µM				Not applicable
	* tail of embryo was damaged during dechorionation (brown spot)			
				

Table 9.5. Continued.

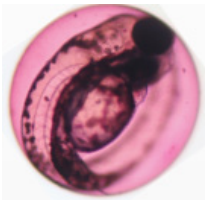
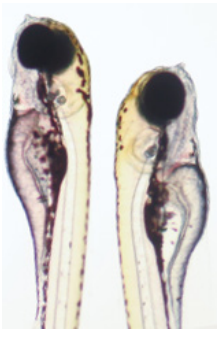
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
10,000 µM				Not applicable
				

Table 9.6 gives an overview of embryos and larvae exposed to Giemsa stain (logP -0.77, stains blue). After 24 hours of exposure no blue colouration of the embryos was visible for both concentrations tested. After 48 hours of exposure to the highest concentration, the ZF embryo appears blue, which is more apparent at 72 hours. However, this effect does not seem apparent anymore at 96 hours of exposure. This could be due to higher pigmentation of the ZF larva at this point, overruling the blue colouration. Moreover, no apparent toxic or teratogenic effects were found after exposure to Giemsa stain, although hatching was delayed when exposed to 500 µM causing a bending of the body.

Table 9.6. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Giemsa stain.
LogP: -0.77. Colour staining: blue.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

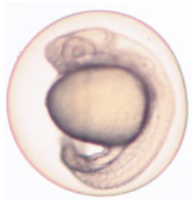
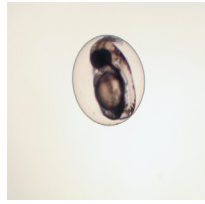
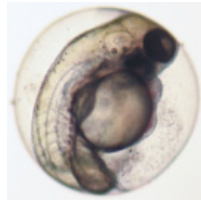
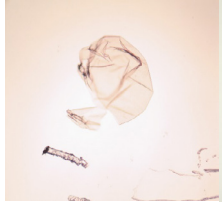
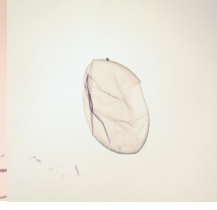
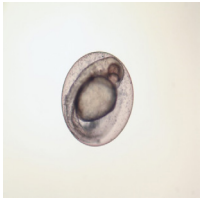
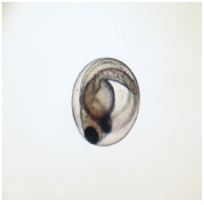
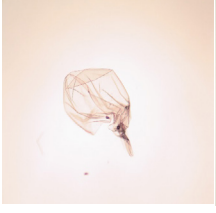
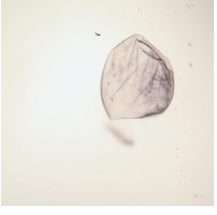
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
250 µM				Not applicable
				

Table 9.6. Continued.

	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
500 µM				Not applicable
				

Visualization of compound uptake in ZF embryos and larvae after exposure to Van Gierson stain (logP 1.64, stains yellow) is depicted in Table 9.7. Although the yolk sac of larvae after 96 hours of exposure appears slightly more yellow than the control in the highest concentration, no clear colouration of embryos or larvae is seen on the other time points. This could be due to the yellow colour of this dye, for ZF embryo's and larvae have a yellowish glow in nature. Moreover, no toxic or teratogenic effects other than a slight oedema of the heart at 96 hrs of exposure to the highest concentration were observed.

Table 9.7. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Van Gierson stain.
LogP: 1.64. Colour staining: yellow.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

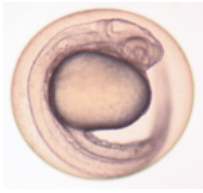
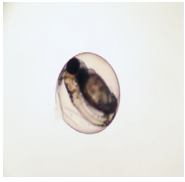
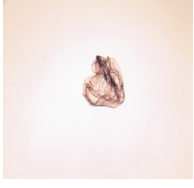
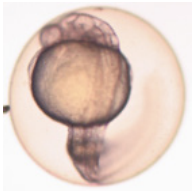
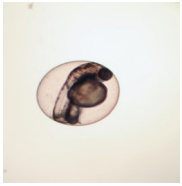
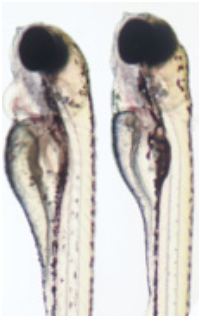
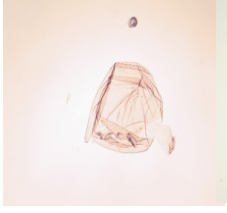
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
5000 µM				Not applicable
				

Table 9.7. Continued.

	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
10,000 µM			Not applicable	Not applicable
				
				

A clear purple colouration, especially within the yolk sac, can be seen in embryos and larvae exposed to Cresyl fast violet (logP 3.5, purple dye), as illustrated in Table 9.8. This effect is even more apparent in the highest concentration tested compared to the lowest, indicating concentration-dependence. Unfortunately, the higher concentration was too toxic for the larvae, resulting in no survival after 96 hours of exposure. Toxic and teratogenic effects were observed after exposure to 10,000 µM, which were bending of the body and delayed hatching.

Table 9.8. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Cresyl fast violet.
LogP: 3.5. Colour staining: purple.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

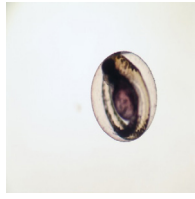
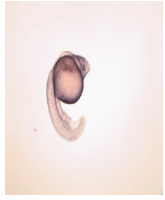
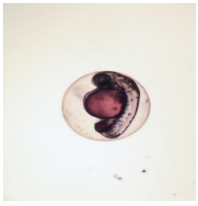
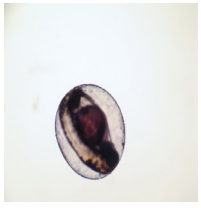
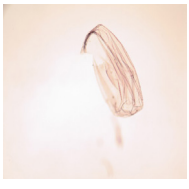
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
5000 µM				Not applicable
				

Table 9.8. Continued.

	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
10,000 µM				Not applicable
				0% survival
				

Although only slight orange colouration is found in embryos and larvae exposed to Eosin Y (solution A) (logP 4.8, orange dye), the eggs appear bright orange (Table 9.9). This is due to orange colouration of the chorion, rather than the PVS, as was apparent during manual dechorionation. An orange colouration of the larva is especially seen after 72 and 96 hours of exposure as an orange colour is found above the yolk sac. Due to toxicity, no larvae survived after 96 hours of exposure at the highest concentration tested. In these embryos/larvae, notable toxic/teratogenic effects were seen at both concentrations, which were oedema of the heart and deformation of the head, with also a retarded growth compared to control embryos/larvae. This indeed indicates compound uptake.

Table 9.9. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Eosin Y, solution A.
LogP: 4.8. Colour staining: bright orange.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

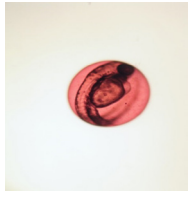
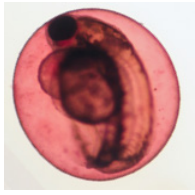
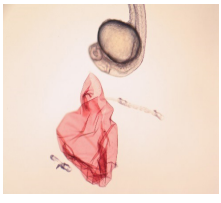
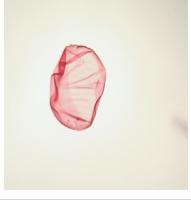
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
250 µM				Not applicable
				

Table 9.9. Continued.

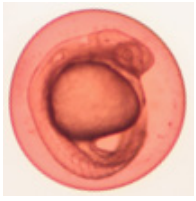
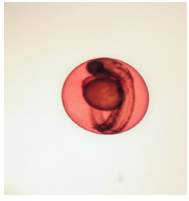
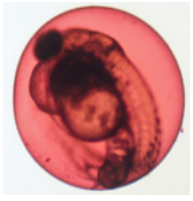
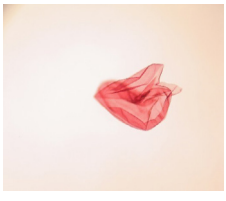
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
500 µM				Not applicable
				0% survival
				

Table 9.10 illustrates the colouration of embryos and larvae after exposure to Sudan III (logP 7.5 red dye). As expected from its high logP value, a stronger red colouration within the whole embryo and larva is seen over time, indicating strong compound uptake. The toxic effect, as apparent from the malformations at both concentrations in the embryos and larvae; oedema of the heart, flat head, small eyes and yaw, also indicates compound uptake. These effects worsened in the higher concentration. Additionally, no larvae survived after 96 hours of exposure to 2000 µM, due to toxicity.

Table 9.10. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Sudan III.
LogP: 7.5. Colour staining: red.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

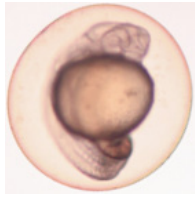
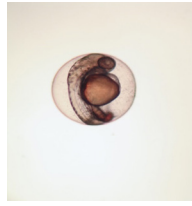
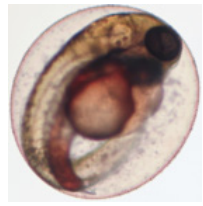
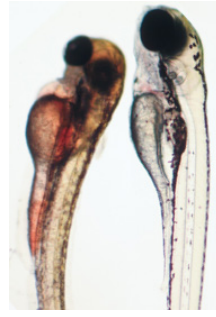
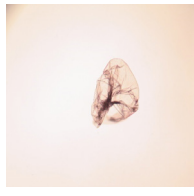
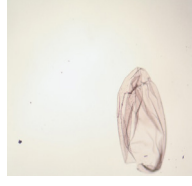
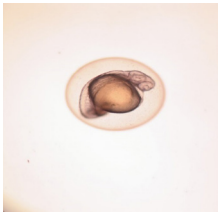
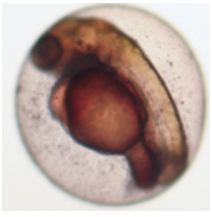
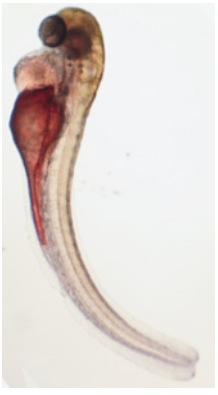
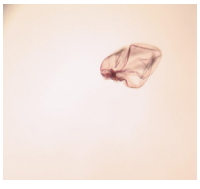
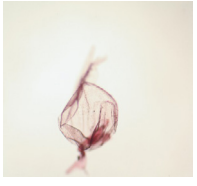
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
1000 µM				Not applicable
				
				

Table 9.10. Continued.

	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
2000 µM				Not applicable
				0% survival
				

The effects of exposure to Oil red O (logP 9.81, red dye) are depicted in Table 9.11. Compound uptake of this dye can be concluded from its toxic effect, as apparent from malformations in these embryos and larvae at both concentration tested; oedema of the heart, bending of the body, retarded yaw and body growth. In this case, compound uptake is not very clear from colouration of the embryo and larva compared to colouration from Sudan III, which has a lower logP value but shows a brighter colouration at the same concentrations tested. This was not due to differences in media colouration, which had the same intensity of colour. However, the embryos and larvae do show red colouration throughout the whole body, especially after 96 hours of exposure, which clearly indicates exposure.

Table 9.11. Visualization of compound uptake and distribution within zebrafish embryos after exposure to Oil red O.
LogP: 9.81. Colour staining: red.
Time points represent hours of exposure. Pictures are taken in ISO-medium, both before and after dechorionation, where applicable.

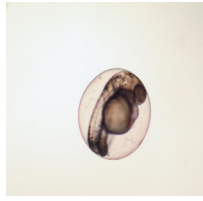
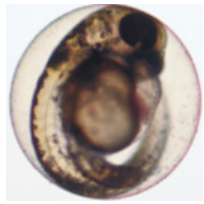
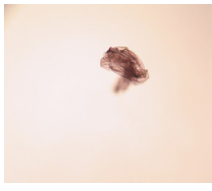
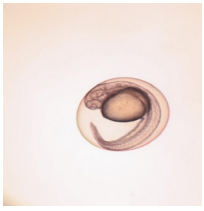
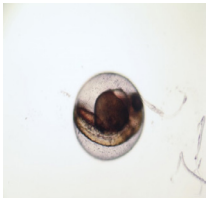
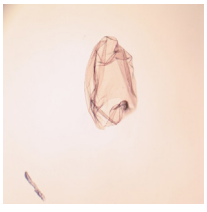
	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
1000 µM				Not applicable
				

Table 9.11. Continued.

	24 hrs	48 hrs	72 hrs	96 hrs (left) versus control (right)
2000 µM				Not applicable
				

9.3.2. Visualization of the effect of DMSO on compound uptake and distribution within the zebrafish larva

Table 9.12 gives an overview of the effect of 1% DMSO on compound uptake and distribution within ZF larvae after 96 hours of exposure to the same dyes as described in the previous section. As no larvae survived until 96 hours with exposure to Giemsa stain, only six dyes are represented in these data. The toxicity found for this concentration is probably due to sensitivity of the batch, as higher concentrations of Giemsa stain, even without DMSO, did not cause these severe toxic effects in Table 9.6. This however, cannot be confirmed as no specific analyses of test concentrations were performed in this study. In all pictures from Table 9.12, the left larva is exposed to the dye with 1% DMSO and the right larva without DMSO. As shown for all dyes, no differences were seen in colouration with and without 1% DMSO. However, if present, it could also be that these were too subtle to be visualised with this method.

Table 9.12. Visualization of the effect of DMSO on compound uptake and distribution within the zebrafish larva after 96 hours of exposure.

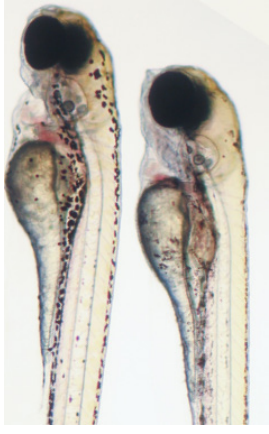
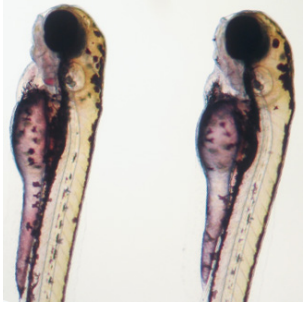
Schiff's reagent LogP: -4.63 Colour: pink 10,000 µM		Left: 1% DMSO Right: no DMSO
Giemsa stain LogP: -0.77 Colour: blue 100 µM	0% survival	Not applicable
Van Gierson stain LogP: 1.84 Colour: yellow 10,000 µM		Left: 1% DMSO Right: no DMSO
Cresyl fast violet LogP: 3.5 Colour: purple 5000 µM		Left: 1% DMSO Right: no DMSO

Table 9.12. Continued.

Eosine Y, solution A LogP: 4.8 Colour: orange 100 µM		Left: 1% DMSO Right: no DMSO
Sudan III LogP: 7.5 Colour: red 1000 µM		Left: 1% DMSO Right: no DMSO
Oil red O LogP: 9.81 Colour: red 1000 µM		Left: 1% DMSO Right: no DMSO

9.3.3. Evaluation of compound uptake within zebrafish embryos and larvae with LC-MS

Figures 9.1 to 9.3 give an overview of the effect of logP on compound uptake and distribution within ZF embryos during exposure to Acyclovir (logP -1.56), Zidovudine (logP 0.05), and Metoprolol Tartrate Salt (logP 1.8). The chorions of embryos and larvae that did not hatch yet were removed prior to measuring internal exposures. The chorion was present throughout the exposure period (i.e. it was only manually removed for bioanalysis). In short, as illustrated by these figures, the internal amount of test compound, in ng per mg tissue in embryos and larvae, increases when the concentration of the compound increased. This indicates a concentration-dependent uptake of test compound by the ZF embryo. Even within the presence of the chorion during exposure, test compound was detected in embryos for all three compounds tested. This shows that the chorion does not act as a full barrier for these compounds but could be rate limiting. However, a slightly higher amount of compound is found in hatched larvae at 72 hrs of exposure, at least for Zidovudine and Metoprolol. At comparable exposure concentrations, the internal amounts found in embryos exposed to Acyclovir (logP -1.56) are lower than the amounts found for Zidovudine (logP 0.05). Also, the amounts found for Zidovudine are lower when compared to the amounts of Metoprolol (logP 1.8). This indicates the internal exposure in embryos and larvae is higher, when logP value is higher.

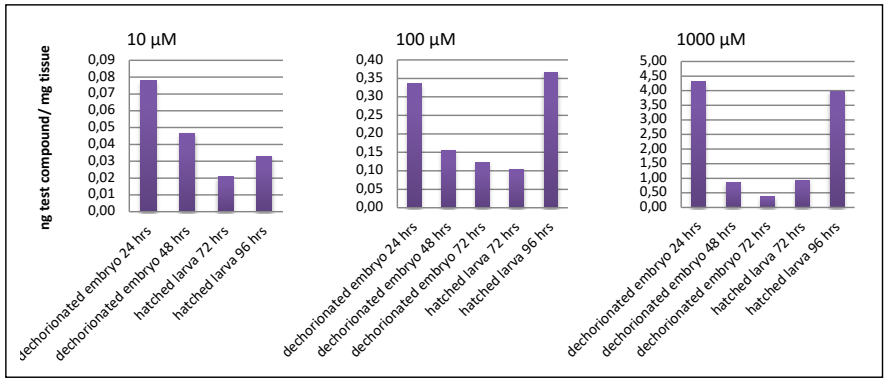


Figure 9.1. Amount of Acyclovir (logP: -1.56) in ng per mg tissue in embryos and larvae. Chorions were present during exposure. Internal exposure after 24, 48, 72, and 96 hrs in embryos and larvae, for 10, 100, and 1000 µM Acyclovir.

For Acyclovir (Figure 9.1) in all concentrations, internal exposure decreases over time in the embryos, with the chorion removed before measuring, until 72 hrs. Internal exposures of 0.1 and 1 µM were below the lower limit of quantification (lloq) and could therefore not be measured. This was also the case for the sample of dechorionated embryos at 72 hrs after exposure to 10 µM. In the highest two concentrations, internal exposure in hatched larvae at 96 hrs increases again. At 72 hrs, no difference between non-hatched and hatched larvae could be found for 100 µM Acyclovir, while for 1000 µM the amount of Acyclovir seems slightly higher after hatching. Furthermore, the factors between the different medium concentrations correspond to the factors between in-

ternal exposure concentrations; a ten times higher medium concentration results in about ten times higher internal exposure, indicating concentration-dependence just as in the results for visualization of compound uptake.

Figure 9.2 shows the amount of Zidovudine in ng per mg embryo and larva, without chorions, after exposure to all tested concentrations. The chorions were present during exposure, before hatching. For the lowest two concentrations of Zidovudine only peaks above the lloq could be found for hatched larvae at 72 and 96 hrs, respectively, and are thus not shown here. In all other concentrations all samples showed quantifiable peaks. Also here, exposure concentrations are medium concentration-dependent. Moreover, in all three highest concentrations, the amount of test compound in the embryo increased notably from 24 hrs to 48 hrs. After this, the amount in larvae at 72 hrs is decreased. The amount of Zidovudine in hatched larvae at 72 hrs further decreases at 96 hrs, indicating the relative uptake of Zidovudine decreases over time. Moreover, no apparent differences can be found in internal concentrations of larvae that did not hatch yet, and hatched larvae, although exposure in hatched larvae at 72 hrs is somewhat higher compared to non-hatched larvae.

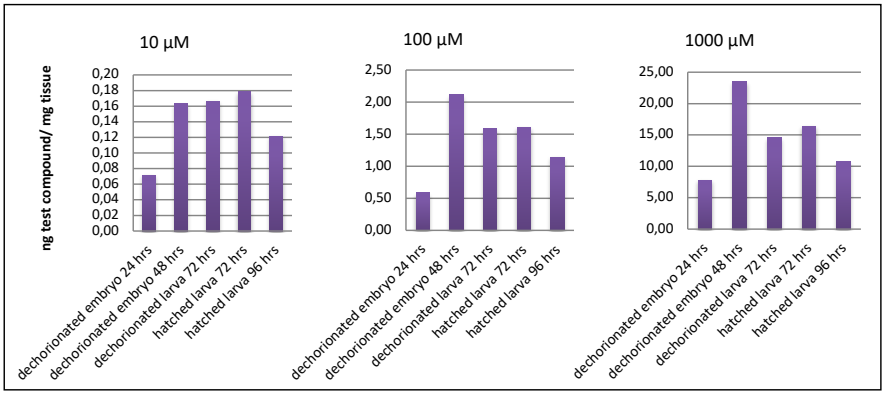


Figure 9.2. Amount of Zidovudine (logP: 0.05) in ng per mg embryos and larvae. Chorions were present during exposure. Internal exposure after 24, 48, 72, and 96 hrs in embryos/larvae and hatched larvae after exposure to 10, 100, and 1000 µM Zidovudine.

Figure 9.3 depicts the concentration of Metoprolol in ng per mg tissue embryo and larva after exposure to 1, 10, 100, and 1000 µM of Metoprolol, while 0.1 µM did not give peaks above the lloq. For Metoprolol internal exposure is concentration-dependent as well. The amount of Metoprolol seems to relatively increase over time in embryos and larvae after exposure to 1 and 10 µM Metoprolol. However, this pattern is not found for 100 and 1000 µM as here the internal exposure stabilizes after a peak at 48 hrs. As the internal exposure is relatively high compared to Acyclovir and Zidovudine, it could well be that the latter observation is due to saturation within the embryo. Moreover, internal amounts of Metoprolol are lower in non-hatched compared to hatched larvae at 72 hours of exposure. This could mean the chorion acted as a barrier for compound uptake.

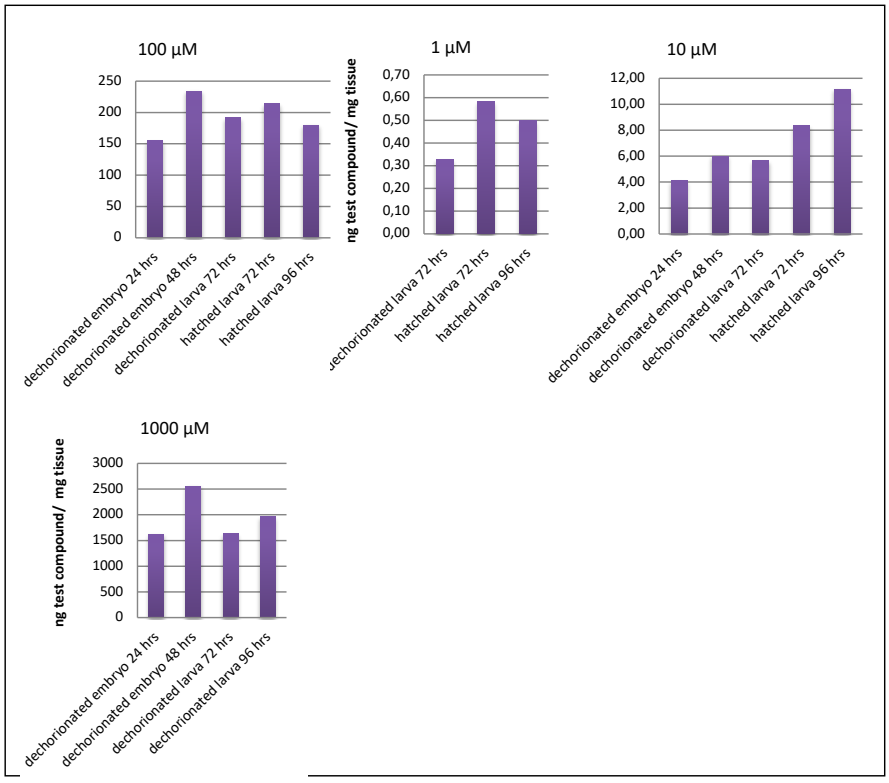
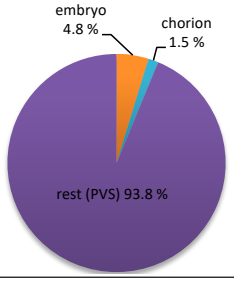
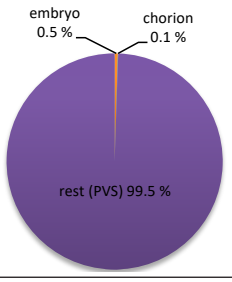
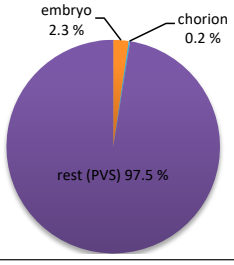
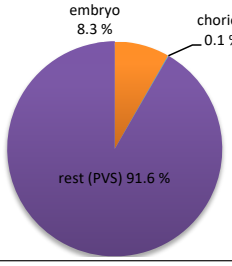
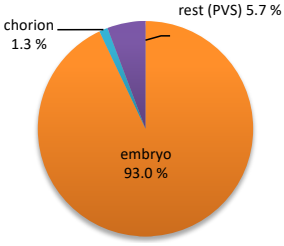
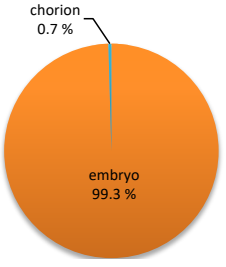


Figure 9.3. Amount of Metoprolol (logP: 1.8) in ng per mg embryos and larvae. Chorions were present during exposure. Internal exposure after 24, 48, 72, and 96 hrs in embryos and larvae, for 1, 10, 100, and 1000 µM Metoprolol.

9.3.4. Distribution of compound within the complete zebrafish egg

The distribution of test compound within the chorion and embryo, as part of the complete egg, after exposure to 1000 µM Acyclovir, Zidovudine, and Metoprolol is depicted in Table 9.13. As can be seen from this overview, for Acyclovir and Zidovudine, both having low logP values, only a small part of the total amount of compound within the complete egg can be found in the embryo and chorion, both in eggs at 24 and 48 hours of exposure. This indicates another factor should be taken into account in order to explain the total amount found within the complete egg, which is the PVS. For Acyclovir, the amount of test compound within the embryo decreases from 24 to 48 hours. However, for Zidovudine an increase is seen over time. This can be explained by the difference in lipophilicity: Acyclovir is hydrophilic, while Zidovudine is slightly lipophilic. For Metoprolol, the most lipophilic compound tested, the vast amount of compound is found within the embryo which increases over time (from 24 to 48 hours). This indicates a different distribution within the ZF egg for lipophilic compounds and hydrophilic compounds which facilitates the finding obtained in the visualization experiment. Overall, test compound did not distribute into the chorion at a notifiable level, as measured in chorions for all three compounds.

Table 9.13. Distribution of compound within the complete ZF egg as percentages.
Percentage of Acyclovir, Zidovudine, and Metoprolol in ng per mg tissue in 24 and 48 hrs eggs, of embryos, chorions, and rest amount after exposure to 1000 µM.

	24 hours of exposure	24 hours of exposure
Acyclovir **logP: - 1.56** 1000 µM 100 % 24 hrs = 91 ng 100 % 48 hrs = 132		
Zidovudine **logP: 0.07** 1000 µM 100 % 24 hrs = 337 ng 100 % 48 hrs = 284 ng		
Metoprolol **logP: 1.8** 1000 µM 100 % 24 hrs = 1755 ng 100 % 48 hrs = 1974 ng		

9.4. DISCUSSION

In this article, we visualised the effect of lipophilicity on compound uptake and distribution within the ZF embryo/larva, using dyes with a broad range of LogP values. Although the sensitivity of this method has not been tested for our purposes, it was shown useful to visualize internal exposure in ZF embryos and larvae as well as distribution. From our results, we conclude that higher logP value (higher lipophilicity) results in higher compound uptake within the ZF embryo and larva but also that the uptake is increases with the applied dose. This conclusion is based on our results from both visual assessment with dyes and the results obtained using LC-MS analyses. From the measurements using LC-MS it could also be concluded that compounds with lower logP values, and thus less lipophilic, will mostly stay in the PVS. Although the chorion did not block compound uptake, it seemed to slightly inhibit this as we found somewhat higher concentrations in hatched larvae compared to non-hatched larvae at 72 hrs of exposure.

9.4.1. Compound logP value visibly affects compound uptake and as well as distribution in zebrafish embryos

Visualization of internal exposure was established for Schiff's staining (logP -4.63) at 10,000 µM (but not 5000 µM), Cresyl fast violet (logP 3.5) at 5000 and 10,000 µM, Eosin Y stain (logP 4.8) at 250 and 500 µM, Sudan III (logP 7.5) at 1000 and 2000 µM, and Oil red O (logP 9.81) at 1000 and 2000 µM. These compounds showed higher colouration of the embryos with higher logP value. However, no evident visible colouration was found for Giemsa staining at 250 and 500 µM (logP -0.77) and Van Gierson stain at 5,000 and 10,000 µM (logP 1.64). As all compounds have a molecular weight much lower than 3000 g/mole [Creton, 2004], this is not caused by lack of uptake due to chorion presence. The lack of colouration found for these two compounds could be due to their structure, inhibiting compound uptake. For Giemsa staining, however, this is contradicted by the malformations found for the embryos and larvae, indicating compound uptake. As Giemsa staining exists of several components the toxicity could be due to uptake of one of its components other than the blue stain. Another probable explanation for the lack of a visible colouration of the ZF embryos and larvae exposed to the blue Giemsa staining and the yellow Van Gierson stain is that these colours are not apparent due to the colour of the embryo. A yellow dye is not easily made visible in a naturally yellow fish. In addition, a blue dye would appear somewhat greenish in a yellow ZF embryo/larva, making it hard to discriminate between the green and the yellow colour. Thus, as the results are not quantitative, it is possible that small amounts of dye, not discernable to the naked eye, could be present in compartments that appear to be devoid of dye.

Interestingly, looking at compound distribution, the colouration after exposure to compounds with a lower logP value (Schiff's staining and Cresyl fast violet) is most apparent within the yolk sac. On the other hand, colouration after exposure to the more lipophilic compounds (Eosin Y (solution A), Sudan III, and Oil red O) distributes throughout the whole embryo. This difference in compound distribution is very interesting, as it emphasizes a new possible explanation for the low toxicity found in ZF embryos exposed to compounds with low logP value, and the high toxicity found with high logP value [Padilla 2012; Steinberg 2013], rather than the amount of internal exposure alone; distribution of compound only within the yolk sac would probably be less toxic compared to discribution of compound within the whole body. However, these issues should be addressed with analytical method in future research, evaluating more compounds with comparable logP values.

Moreover, a very apparent colouration of the chorion was found for Schiff's staining and Eosin Y, (solution A) stain. Here, the embryos and larvae showed minor colouration, while the chorions appeared brightly coloured. As the logP values of these compounds are very different (logP of -4.63, and 4.8, respectively), this is not due to the level of lipophilicity. However, this could be due to high compound uptake by the chorion, or accumulation of compound molecules on the surface of the chorion. As shown in the LC-MS results, however, the chorion likely only accounts for a small proportion of compound uptake at least for the compounds tested here. So it is highly unlikely that these dyes are actually taken up by the chorion. Either way, it is not clear whether this affected internal exposure to the embryos. It would therefore be interesting to compare uptake of these compounds in embryos exposed within the presence of the chorion to embryos exposed after dechorionation.

However, when looking at the overall results from exposure to dyes, as the compounds that established visible results represent both low and high logP values, we were able to visualize compound uptake for the more lipophilic compound, as expected, but also for hydrophilic compounds. But, to establish a visible result for the more lipophilic compounds such as for Schiff's staining, which has the lowest logP value, a concentration of 10,000 µM was needed. For compounds with higher logP values, a lower concentration already showed colouration in the ZF embryos and larvae. This indicates a higher logP value results in visible compound uptake at lower concentrations compared to lower logP values. Thus internal exposure for compounds with a higher logP value is higher than for compounds with a lower logP value. The same can be concluded based on the results of the concentration measurements by bioanalysis.

9.4.2. Specific analysis supports the finding that logP affects zebrafish embryo compound uptake as well as distribution

The finding that internal exposure of ZF embryos/larvae can be related to logP value was also found by Berghmans and colleagues, who found a higher internal concentration of test compounds with higher logP values as part of their internal exposure controls, measured at 96 hrs of exposure [Berghmans, 2008]. Unfortunately, this group did not test compounds with a negative logP value, but they did assess compounds with logP's just above 0. For these compounds they found uptake within the embryos and larvae. So, as suggested by our results as well, a low logP value does not mean there is no compound uptake by the ZF embryo. Moreover, the uptake increases with the applied dose.

Analytical results for Acyclovir (logP -1.56) indicated that internal exposure of Acyclovir in embryos (24 to 48 hpf) decreases in time until 72 hrs. But when hatched at 72 hrs, internal exposure increases again greatly, although internal concentrations of hatched and unhatched larvae are comparable. This could be due to areas of increased permeability because of development of gills and digestive system, which is between 48 and 72 hpf [Kimmel, 1995; Hill, 2003]. For Zidovudine (logP 0.07), the amount in embryos and larvae increased from 24 to 48 hrs. After this time point the amount of Zidovudine decreased until 72 hrs in the higher concentrations, but stayed comparable to the 48 hrs-level when exposed to 10 µM. For Metoprolol (logP 1.8), the amount of internal exposure in embryos seems to increase until 48 hrs of exposure, after which it was noted that exposure stabilizes for the lower concentrations, or rises for the higher concentrations, indicating saturation. Because no apparent increase was found after 72 hrs for Zidovudine and Metoprolol, no effect from development of gills and the digestive system on compound uptake was indicated for these compounds. Moreover, the amount of Zidovudine in hatched larvae at 96 hrs was lower than after 72 hrs of exposure. The latter could be due to a higher growth rate of the larva when compared to compound uptake rate, causing a lower amount of test compound per mg tissue.

Importantly, as the chorion is present until hatching after 72 hrs of exposure and compound uptake is found already at 24 and 48 hrs, the chorion does not inhibit compound uptake by the embryo. It could, however, limit the amount of compound taken up. This could especially be the

case for compounds with a logP value of 0 or higher, as for Zidovudine and Metoprolol we found that amounts in non-hatched 72 hrs old larvae were slightly lower compared to hatched larvae at 72 hrs. This thus needs to be investigated by comparing internal exposures in dechorionated and non-dechorionated embryos.

When looking at the distribution of compound within the embryo and chorion as part of the complete egg, we saw that a vast part of total Acyclovir and Zidovudine is not present in the embryo or the chorion. This indicates that this amount of test compounds resided in another part of the egg, which is the PVS. As Acyclovir and Zidovudine have low logP values, this can be expected as the embryo is a lipid structure. However, it is indicated here that for compounds with lower logP value and thus more hydrophilic, the PVS might influence exposure to the embryo in the presence of a chorion [Brox, 2014]. For Metoprolol nearly all of the compound was found within the embryo. This is expected as Metoprolol is a lipophilic compound (logP 1.8) and therefore likely to reside within the more lipid embryo tissue. Thus for lipophilic compounds, the PVS does not influence exposure while its effect on exposure to hydrophilic compounds should be investigated further. It should be noted here, that analysis of internal exposure in the complete egg can thus be misleading.

9.4.3. The use of DMSO as a solvent does not have any visible effect on compound uptake and distribution within the zebrafish embryo

Using visualization of compound uptake and distribution within the ZF embryo and larva, we were able to make an easy comparison between exposure with or without 1% DMSO in the medium. Although some groups suggested that DMSO affects compound uptake in ZF embryos and larvae [Hallare, 2006; David, 2012; Kais, 2013], we could not visualize a difference in colouration between ZF larvae exposed to dyes with or without additional DMSO for all compounds tested. If indeed DMSO influences the uptake of substances, this influence was too subtle to be visualised with this method.

9.4.4. When does high uptake cause false positive results and low uptake false negative ones?

Overall, our results from both visualization and bioanalysis showed that internal exposure for compounds with low logP values was much lower than for compounds with high logP values, which can be explained by the lipid structure of the ZF embryo/larva. For hydrophilic compounds, this could easily cause false negative results. On the other hand, more false positive results will be caused by higher internal exposure for the more lipophilic compounds, due to compound overload [Bulck, 2011]. Therefore, it will be of utmost importance to define these threshold values (both minimum and maximum).

For this, Gustafson and colleagues set a threshold at 50 ng/larva, above which internal exposure would cause false positive results [Gustafson 2008]. On the other hand, Ball and colleagues came with a classification in percentage of compound uptake, as compared to the exposure medium, in which internal exposure at 24 hrs below 5 % is classified as low, and therefore causes false negative

results [Ball, 2014]. However, when Metoprolol is taken as an example for the threshold set by Gustafson, the mean body burden found for 1000 µM Metoprolol (460 ng/larva) is above 50 ng/larva and therefore would give false positive results. On the other hand, based on the statement of Ball and colleagues, the percentage of uptake after 24 hrs at 1000 µM Metoprolol is 1.6 %, which is below the set value of 5 %. This would classify the internal exposure of 1000 µM Metoprolol as low, making it likely to give false negative results. From the finding that the mean body burden would classify exposure to 1000 µM Metoprolol as false positive, while the percentage of uptake would classify 1000 µM Metoprolol as false negative, it is questionnable which indication is reliable for evaluation of compound uptake. This indicates a high need to search for true thresholds for internal exposure, to adres the problem of false positive and false negative results. The latter also stresses the importance of bioanalysis, when testing with ZF embryos, as the logP value of a compound highly affects internal exposure and distribution.

9.5. CONCLUDING REMARKS

As compound uptake and thus internal exposure of the ZF embryo highly influences test results, it is important to know what factors are of influence on this. For this reason, we evaluated two factors that were thought to highly affect compound uptake in ZF embryos, i.e. logP value and the use of DMSO as a solvent. In order to actually visualize the effects of these factors, the use of dyes, for colours that are distinctive from the natural ZF larva colours, showed to be very practical. In addition, this method is found to be reliable as the results from visualization are in line with our results from bioanalyses, both showing that higher logP value results in higher internal exposure. Also, the chorion did not seem to function as a barrier for compound uptake, but might be rate limiting.

Furthermore, we were the first to measure the actual distribution of test compounds within the ZF egg every 24 hours, taking the amount found in the embryo and the chorion as separate parts of the complete egg. Here, using LC-MS, we showed that for lipophilic compounds, >90% of compound within the complete egg is present within the embryo. For hydrophilic compounds, >90% of compound was probably distributed into the PVS. Therefore, the precise effects of the PVS, but also the chorion, should be clarified in further research.

We showed that logP value not only highly influences compound uptake in the complete egg, but we also found a striking difference between compound distribution within ZF embryos between low and high logP values. Furthermore, next to logP value, factors like molecular weight, molecular charge, and molecular structure could also highly influence uptake and thus internal exposure within the ZF embryo. Therefore, to measure and quantify internal exposure, it is detrimental to apply bioanalysis additional to tests done with ZF embryos. This should be done on at least two occasions (i.e. before and after hatching). Moreover, before hatching the egg should be manually dechorionated to determine the concentration within the embryo as opposed to the complete egg.

Chapter 10

Summary, general discussion, future perspectives and conclusion

Adapted from:

Manon Beekhuijzen, The era of 3Rs implementation in Developmental and Reproductive Toxicity (DART) Testing: current overview and future perspectives, Reproductive Toxicology, 2017 Sep;72:86-96. doi: 10.1016/j.reprotox.2017.05.006. Epub 2017 May 25.

10.1. SUMMARY OF MAIN FINDINGS AND GENERAL DISCUSSION

The aims of this thesis were:

- to review and critically assess, to practically improve and to test methods of (recently updated) regulatory guidelines for developmental and reproductive toxicity (DART) testing in respect of the 3R principles (i.e. reduce, refine and replace), and
- to provide an outlook towards future advances in DART testing.

Within the setting of a Contract Research Organization (CRO) the availability of data from numerous projects produced by highly trained technicians enabled the optimal environment to perform the work described in this thesis. These papers can be used as guidance for other laboratories, study monitors and registration officers. Moreover, this work was published to encourage others to discuss and publish their procedures and results, to improve and harmonize these methods worldwide with special focus on refinement, reduction and replacement.

This thesis consists of three parts:

1) The implementation of refinement, and consequently reduction, in current DART studies,
2) The impact on refinement and reduction due to the recently updated/implemented OECD guidelines, and
3) An outlook on the zebrafish embryotoxicity test (ZET) as possible replacement of rats and/or rabbits in future developmental toxicity testing.

10.1.1. The implementation of refinement, and consequently reduction, in current DART studies

The first part of this thesis discusses the implementation of refinement, and consequently reduction, in current DART studies. Refinement of blood sampling techniques for toxicokinetic (TK) assessment of pharmaceuticals greatly improves the animals' well-being by reducing stress, and provide the opportunity to reduce the number of animals that are being used. This was shown by the implementation of capillary microsampling (CMS) in studies with juvenile rats (Chapter 2) and jugular vein blood sampling in rat embryo-fetal development (EFD) studies (Chapter 3).

TK sampling is a necessary aspect of the toxicity evaluation to ensure that appropriate plasma drug levels are reached and that any observed toxicological response is biologically relevant to systemic exposure [Harstad, 2016]. The ICH S3A guideline from 1994 requires TK assessment in toxicity studies to determine systemic exposure as a basis for risk assessment and proper dose selection in the clinic; however, the guidelines do not require the use of satellite animals, only that the minimum number of animals is used to obtain consistent and adequate data [ICH S3A, 1994]. This guideline is undergoing a revision that reflects modern microsampling advances and encourages the reduction or elimination of the number of rodent satellite groups [S3A Implementation Working Group, 2016] as advances in bioanalytical sensitivity have enabled smaller blood collection requirements [Harstad, 2016].

Chapter 2 showed that, after comparing submandibular, saphenous, tail, and jugular vein blood sampling in juvenile animals, the best location for CMS (based on feasibility) was the submandibular vein for postnatal day (PND) 4 and 10 pups and the tail vein for PND 17 pups. Moreover, PND 4 pups can be sampled twice and PND 10 or 17 pups can be sampled on three occasions within 24 hours. Strikingly, by using this method it can decrease the number of satellite juvenile rats by 75%. This is based on the following calculation: in a traditional study, two pups are sacrificed to collect one TK sample at each time point. Normally 4 dose levels and 6 time points are included in a preclinical rat toxicity study, with 3 animals/sex/group. Thus, 288 satellite PND 4 pups are sacrificed to collect the TK blood samples. Our work showed that using CMS, two samples (instead of ½ sample) can be collected from one PND 4 pup in 24 hours, which would reduce the satellite pup numbers to 72 i.e. by 75%. Noteworthy, as an acute effect of CMS on several hematology parameters (e.g. reduced level of hemoglobin and red blood cells) was seen, it is not suggested to use the main study animals for TK blood sampling but use less satellite animals instead.

There is extensive literature on the use of CMS in adult rats, but the only data using CMS in juvenile rats is a paper on PND 19 pups [Powles-Glover, 2014a] and a poster on PND 9, 14 and 21 [Clubb, 2015]. Since PND 4 is in general the youngest age for dosing the rats in juvenile studies, we started our assessment with PND 4 pups and added PND 10 and 17 to this to cover the pre-weaning period. We were the first to publish work on CMS at such a young age.

CMS is gaining more attention but is not yet used as default. Hurdles in implementation in toxicology studies consist of the following three main categories: 1) toxicological interpretation, 2) bioanalytical and technical feasibility, and 3) regulatory acceptance [Chapman, 2014a; Chapman, 2014b].

The toxicological impact of microsampling has been focused largely on affected hematological parameters and exacerbation of tolerability. Blood volume limitations are commonly applied at 15% of circulating blood volume, largely established to reduce the impact of these effects on animal welfare [Diehl, 2001]. Higher conventional sample volumes may be restricted by TK blood volume needs when using a serial collection schedule [Sparrow, 2011]. When taking this into account, it is still shown (as expected) that levels of hematocrit, red blood cell count and hemoglobin are decreased, and number of reticulocytes are increased when serial TK samples are collected prior to hematology samples [Caron, 2015; Powles-Glover, 2014a; Powles-Glover, 2014b]. This effect is evident with serial blood sampling at higher volumes (6 × 200 µL), whereas repeated smaller volumes (up to 6 × 100 µL) have minimal to no effect [Caron, 2015; Jonsson, 2012a, Powles-Glover, 2014b]. Microsampling techniques have included processes to handle either wet or dry samples along with a range of biological matrices, including whole blood, serum, or plasma, but could be expanded to just about any sample type [Harstad, 2016]. All of these approaches require changes either to sample collection techniques, animal handling, bioanalytical sample preparation, and/or study designs. For many, there is no single best way or best technique as long as the final choice is technically feasible, reproducible, and adheres to ethical standards of animal welfare. In theory, there

is almost no lower limit in volume requirements for microsampling. However, there are logistical concerns for various groups that usually define minimum or maximum volumes that can be taken or worked with conveniently. The sampling site together with the sampling equipment will determine the minimum amount that can be sampled. Separation of blood into plasma and the transfer of plasma into new tubes are extra challenges for efficiently analyzing extremely small volumes. Most people prefer working in volumes ≥ 150 µL, which allows the preferred use of labelled tubes to centrifuge and a sufficient volume (~75 µL) of plasma to be quickly dispensed into a new tube and keeping it feasible [Harstad, 2016].

Perceived regulatory barriers have been suggested by several authors, and indeed little has been published about the extent of successful implementation in regulatory toxicity studies. Chapman et al. [2014b] report that microsampling is commonly adopted in PK studies but not in GLP studies as of 2014. Several reports from industry-leading companies highlight strategies whereby they eliminate satellite TK groups in GLP studies, although the extent of successes and failures has not been published [Harstad, 2016].

In cases when CMS cannot be used, a quick win for enhancement of blood sampling technique is by using the jugular vein. This method gives minimal stress and does not need any anesthesia. However, a high degree of competence is required for this technique and sufficient time should be allowed to fully train the technicians. Correct restraint of the animal and precise insertion of the needle is crucial for successful blood sampling and minimal animal discomfort. We investigated if blood collection from the jugular vein would not affect health and/or embryo-fetal development, and therefore satellite animals would no longer be needed in EFD rat studies. As default, no hematology assessment is needed in these types of studies. **Chapter 3** showed successful implementation of jugular vein blood sampling in an EFD rat study excluding satellite animals, thereby reducing the number of rats in standard EFD studies for pharmaceuticals by 20% (i.e. 21 F0-females and approximately 252 F1-fetuses; assuming a litter size of 12 fetuses/litter). These numbers are based on 6 satellite F0-females per treatment group (as each female would be bled at 3 time points) and 3 satellite F0-females for the control group (as these will only be bled around the anticipated Tmax for the test item groups; 2 time points). This method of blood sampling would therefore be recommend in case microsampling cannot be used. Moreover, similar to microsampling, it may decrease variability due to calculation of TK parameters from composite samples, and would allow direct comparison of individual effects to individual exposures.

One aspect that should be emphasized, is the impact of excluding the use of satellite animals (both with CMS and jugular vein blood sampling) on the assessment of fetal exposure. In case fetal exposure data is requested, maternal animals should be dosed beyond the standard dosing period (until Day 20 post-coitum instead of Day 17 post-coitum). Extension of the dosing period would omit the standard post-dosing recovery period of 3 days which could result in higher toxicity and a lower NOAEL. In rare cases it might underestimate the incidence of fetal malformations as more malformed fetuses may die with prolonged maternal dosing, and resorption rates could be incre-

ased. Fetal malformation rates will not increase because these are induced during the period of major organogenesis (i.e. from Days 6-17 post-coitum), included in both dosing regimens, except for late-induced malformations (e.g. premature closure of the ductus arteriosus). On the other hand, by omitting the treatment-free recovery period it would be a better model for human exposure which is of course the ultimate reason for performing the EFD study. Since numerous organ systems continue to develop in the late phase of the fetal period (in the rat and human), continued exposure during this stage would enable a complete assessment of developmental toxicity.

10.1.2. The impact on refinement of methods and reduction of animal use due to the recently updated/implemented OECD guidelines

The second part of this thesis (**Chapters 4, 5, 6 and 7**) describes the impact on refinement of methods and reduction of animal use due to the recently updated/implemented OECD guidelines 421, 422 and 443 for reproductive toxicity testing. These OECD DART guidelines were updated with additional endpoints necessary for improved human risk assessment. This includes addition of endocrine endpoints by default to all three guidelines, and addition of developmental immunotoxicity and/or neurotoxicity parameters in special cases to the EOGRTS (Chapters 4-7), and is based on increasing incidences of adverse outcomes associated with these endpoints [Kortenkamp, 2011].

Before OECD guidelines 421 and 422 were updated in 2015, we assessed the efficacy of the guidelines dated 1995 or 1996 by performing a retrospective evaluation of 134 reproduction/developmental screening studies (**Chapter 4**). The major findings were:

(1) For up to half of the studies with developmental and reproductive toxicity, these effects would have been missed in other types of studies, which underscores that reproduction/developmental screening studies should not be waived by default based on negative 28-day and/or prenatal developmental data. These effects consisted of reduced numbers of corpora lutea and implantation sites, decreased fertility, conception and gestation indices, increased precoital time, and/or compromised maternal care. These findings would only be detected in reproduction studies.

(2) The required number of animals as stated in the guidelines is appropriate for detecting developmental and reproductive toxicity. In all 27 repro screening studies that showed developmental and/or reproductive effects, the study design resulted in clear and conclusive treatment-related DART effects as an average of 79% of the animals or litters were affected. This shows that, when dose levels are correctly chosen, the compounds that cause developmental and reproductive toxicity show these effects in the majority of animals and/or litters.

(3) Adding measurements like anogenital distance, internal sex determination and nipple retention, plus extending the postnatal period would add predictive value. The importance of an extended postnatal period was assessed using 19 generation studies performed at the same testing

facility. Out of the 8 generation studies with DART findings, 5 of them showed affected pup body weights only, which all occurred after Day 4 of lactation. This was a strong indicator that prolongation of the postnatal period would be crucial for detecting late postnatal effects (i.e. effects potentially induced or manifested after postnatal day 4). Translating these results to repro screening studies, it would be very simple to extend the postnatal period in these repro screening studies, which would provide substantially more data on postnatal development but will not increase the number of experimental animals used.

It was concluded that the current reproduction/developmental screening studies are effective in providing unique data, especially considering the limited number of animals used, but with some simple additions its value in risk assessment could be enriched even further.

The importance of our recommendation on additional measurements and extension of the postnatal period was recognized by the OECD working group and consequently incorporated in the OECD 421 and 422 guidelines update of 2015.

Chapters 5 and 6 provide practical considerations, details on legislation and guidance for implementation of the updated OECD 421 and 422 and newly adopted OECD 443. These chapters were issued to be in line with the 3R principles by sharing our knowledge and provide guidance to other laboratories that are implementing these challenging study designs.

Moreover, this work was published to encourage others to discuss and publish their procedures and results, to improve and harmonize these methods worldwide with special focus on refinement, reduction and replacement.

One item that was critically appraised in **Chapter 5** comprised blood sampling of pups on PND 4. In the 2015 updated OECD 421 and 422 guidelines the following was mentioned "W*hen culling is used, blood samples should be collected from two of the surplus pups, pooled, and used for possible determination of serum T4 levels. If there are insufficient pups in a litter to have two surplus pups, two of the pups that are typically retained would be used. When litter size drops below the culling target (8 or 10 pups/litter), preference will be given to remove two female pups for PND 4 blood samples in order to retain more male pups for nipple retention on PND 13; however, retained female pups in each litter should not drop below 2 if this can be avoided. If litter size is not adjusted, preferably the two pups per litter selected for T4 blood sampling should be female pups to reserve male pups for nipple retention evaluations*". Our paper mentioned several arguments against this procedure, which persuaded the OECD expert group to modify this in the OECD 421 and 422 guidelines. These changes were adopted in 2016. However, not all proposed changes were incorporated. For example, it is still optional to perform histopathological examination of the thyroid according to the 2016 OECD 421 guideline.

The assessment of thyroid hormones, which was recently added for endocrine evalution, has many uncertainties. There is an imperative need for clarification and guidance regarding the collection, assessment, and interpretation of thyroid hormone data for regulatory toxicology and risk assessment, particularly in regard to study designs that include thyroid hormone evaluation of young animals. There are currently cross-laboratory discrepancies in the methods being used to measure thyroid hormones, as well as in the success of obtaining reliable data. Even though publicly available regulatory guidelines and guidance address study design, they present varied approaches to thyroid hormone measurement in young rodents, and an optimal study design or logical approach to thyroid hormone testing in these cases has not yet been established. When apparent treatment-related effects in thyroid hormone levels are observed, questions often arise regarding how much of a change, in what endpoints, should be interpreted as an adverse outcome with potential developmental consequences. Therefore, more guidance on this topic will be crucial in the near future.

The addition of endocrine endpoints to the recently updated OECD 421 and 422 guidelines is considered crucial for gaining more information on the endocrine disruptor potencies of industrial chemicals as mentioned in **Chapter 5**. However, it is important to realize that these additions cause extra animal usage and discomfort. In that respect it should be critically assessed if these changes are worth these consequences, and when sufficient data becomes available from these updated studies, retrospective analyses are needed to draw any conclusion on the additional value of these endpoints versus the impact on the 3Rs.

In **Chapter 6** the hallmarks of the EOGRTS were evaluated and specific considerations that should be taken into account when considering to implement this study design were addressed. It was concluded that the complexity of the study should not be underestimated and experienced testing laboratories with sufficient resources and historical control data for all parameters should conduct the studies. For the coming years, it will be important to gather and review data from the EOGRTS to evaluate the utility of this design for identifying developmental and reproductive toxicants and to further optimize the design if needed. Altogether, this should then also provide more insights on any benefit of an (internal) triggered second generation mating.

Chapters 6 and 7 provide guidance and background on the design of the EOGRTS. The EOGRTS improves upon several aspects of the Two-Generation study by including evaluation of more endpoints. The potential use of triggering of the various cohorts further enhances the flexibility and relevance of the study to assess endpoints of concern while minimizing the use of animals to the fullest extent possible. However, the acceptability and utility of triggering, and the ability for the cohorts to provide sufficient information as stand-alone studies still remain the subject of debate. Taken together, the EOGRTS is a definite step in the right direction of an enhanced study design but with the considerable technical, scientific and regulatory challenges it presents, the ultimate design of the ideal study remains a work in progress.

OECD 443 (EOGRTS) describes three cohorts of F_1 generation animals that may be used to address specific developmental endpoints: reproduction, developmental neurotoxicity and developmental immunotoxicity. The conduct of these cohort evaluations are not mandatory to the validity of the test, nor are they default requirements in all regulatory frameworks, such as REACH (EU 2015/282, 2015). Thus OECD TG 443 offers flexibility in the conduct of these evaluations: "*Decisions on whether to assess the second generation and to omit the developmental neurotoxicity (DNT) cohort and/or developmental immunotoxicity (DIT) cohort should reflect existing knowledge for the chemical being evaluated, as well as the needs of various regulatory authorities*." (OECD 443). In order to reduce the number of animals and procedures involved in meeting the requirements of specific regulatory frameworks, a set of triggers and waivers for the inclusion or exclusion of evaluations was proposed that allow the EOGRTS to not only reduce animal use, but to refine their use in terms of limiting unnecessary manipulations. This is a first step in dedicated case by case testing based on data of the test compound instead of performing a standard list of studies for all test compounds, which will enhance the relevance of testing.

The major challenge is how to decide on triggering of the modular aspects of the study on the basis of prior knowledge of a particular chemical, or on earlier findings in the EOGRTS itself. **Chapter 7** offers guidance on science-based triggers for these extended evaluations. Details on triggers for all cohorts are also described in the ECHA guidance document on information requirements and chemical safety assessment [ECHA R.7a, 2016]. This guidance states that the study design should be decided before the study is started and does not recommend the use of in-study (''internal'') triggers. In addition, Column 2 in the REACH Annexes specifies the expansion criteria to the basic configuration of the EOGRTS. In practice, performance of a cohort will be requested as soon as doubts arise.

In case of equivocal reproductive and/or developmental findings in the first generation offspring, mating for a second generation is warranted. However, in case clear reproductive and/or developmental findings are obtained in the first generation offspring, we would not recommend production of a second generation, as in general, this would not be of any added value and it would increase the number of animals substantially (**Chapter 6**). Moreover, several retrospective studies concluded that a second generation will not significantly change the risk assessment or classification and labelling. It would only confirm results already seen and would rarely, if ever, provide significant new information leading to adjustment of regulatory decisions, e.g. setting the overall NOAEL for risk assessment or discerning between reproduction toxicity or developmental toxicity Category 2 (i.e. "suspected of damaging") or Category 1B ("may damage fertility or the unborn child") for classification and labelling [Janer 2007; Martin 2009; Piersma 2011; Rorije 2011]. This was reported as OECD document no. 176 [OECD no. 176, 2012].

The ECHA guidance document states "*So called internal triggers or in-study triggers for mating the Cohort 1B animals to produce the F2 generation (as those described in OECD GD 117) are not recommended to be used as such in REACH. However, the registrant may expand the study based*

on new information indicating a concern which needs to be addressed." It should therefore be realized that in case the EOGRTS is being performed for other registration purposes than REACH, a different approach for the study design applies (e.g. guidance document 117 is only applicable for the United States and Canada [OECD GD 117, 2011]).

Based on the different regulatory demands, the flexible design of the EOGRTS poses some challenges in case a compound needs to be registered under several international regulations:

- Under REACH it might not be needed to include the DNT and/or DIT cohorts, but this is included by default for US EPA registries.
- The US EPA agrees with the use of internal triggers (for inclusion of a second generation). This is however not allowed under REACH.
- For Japan, a two-generation study is still requested. Which currently means that companies are performing an EOGRTS including the second generation even though it might not be requested for Europe or the US, and thereby using substantially more animals.

It would be important to align these regulatory requirements so all authorities agree on one approach.

Based on present knowledge it can be argued that a second generation should always be included in the case of equivocal results from the parental or first generation. However, there is no added value in extending to a second generation when already clear reproductive and/or developmental toxicity is noted during the first generation. This would only confirm results already seen and would rarely, if ever, provide significant new information leading to change in regulatory decisions. Moreover, it should be realized that the decision to mate by using internal triggers has three inherent challenges requiring 1) planning and dedicating resources for rooms and personnel, 2) a rapid and thorough interpretation of the data, and 3) a review and discussion by involved parties to expedite a consensus. All results for making the decision based on internal triggers become available at a very late stage (particularly P reproductive organ histopathology and F_1 estrous cycle data). This leaves only a limited period for review and discussions prior to making a very important final decision.

OECD guidance no. 117 [OECD GD 117, 2011] mentions all potential effects which may be used as internal triggers; however, the final decision may be subjective and challenging. For instance, the guidance document states that F_1 animals should be mated in the case of biologically relevant decreased F_1 pup body weights in the absence of maternal body weight decrements. What criteria should be applied to demonstrate biological relevance? Should there be statistical significance? What if only a few litters are affected? These considerations should be established on a case-by-case basis and evaluated in great detail. It is therefore of utmost importance that only experienced study directors with a strong scientific support team are appointed because once it is decided not to continue, there is no possibility to later rectify this. Moreover, in case dose level selection is

performed according to OECD 443 i.e. some maternal toxicity at the high dose level, in most occasions this will also lead to some general developmental toxicity i.e. slightly lower body weight of the pups. Would this automatically trigger mating for a second generation?

It is important to realize that for all EOGRTS in which second generation mating is not included beforehand, it is necessary to schedule room and personnel for a possible extension to a second generation as this could be needed in case of equivocal findings. Fortunately, as this decision is based on data of the lactation period of the F_1-generation, there is more time available for decision-making.

Overall, it can be concluded that the approach that ECHA is currently using i.e. based on available data of the test compound a decision is made on the requirement of the second generation, developmental neurotoxicity cohort and developmental immunotoxicity cohort, is in respect of the 3Rs very valuable. As this is a first step towards more tailor-made approaches for regulatory safety assessment instead of tick-box driven standard animal experiments.

10.1.3. Possible replacement of animals in future developmental toxicity testing

The third part of this thesis (**Chapters 8 and 9**) provides an outlook on the zebrafish embryotoxicity test (ZET) as possible replacement of rats and/or rabbits in future developmental toxicity testing. The zebrafish (ZF) has emerged as a promising alternative animal model to screen for teratogenicity because it is cheaper and faster, when compared to tests with mammals. Moreover, it complies with the 3R's framework as the ZF model is not regulated by current legislations on animal welfare up to 120 hours post fertilization (hpf). Not only the rapid organogenesis of the transparent ZF embryo, but also the simplicity of maintaining and breeding the adult ZF, makes this animal model ideal for developmental toxicity screening.

However, the methodology of the ZET used by the different groups varies greatly and harmonization of procedures is crucial. **Chapter 8** describes the most optimal study design that was based on literature and empirical data. Based on literature and/or guideline experience, some conditions were already described in detail and therefore the most optimal test condition could be proposed with confidence (e.g. choice of concentrations, exposure method, and use of solvent). But for other parameters, very limited information was available to definitely propose the most optimal ZET condition (e.g. temperature, and threshold value for teratogenicity). Therefore, it is of great importance that first characterization of the test conditions is needed, followed by experimental validation by different groups on a wide variety of compound classes to confirm its optimal prediction, before harmonization can be achieved.

In 2009, the Health and Environmental Sciences Institute (HESI) of the International Life Sciences Institute formed a steering committee to revisit the issue of a consensus list of developmental toxicants. This committee concluded that rather than agreeing on a list of developmental toxic

chemicals, a list of developmental toxicant exposures should be developed [Lee 2012]. Daston et al [Daston, 2014] selected a series of 20 chemicals that caused developmental toxicity and for which there were appropriate toxicokinetic data. For each chemical, this list states a positive exposure (i.e. the Cmax at a dose level that produced significant teratogenicity or embryolethality) and negative exposure (i.e. the Cmax at a dose level not causing developmental toxicity). This exposure-based reference list represents a fundamentally different approach to the evaluation of alternative tests and is proposed as a step toward application of alternative tests in quantitative risk assessment. Currently, this list is being tested by several laboratories worldwide to validate its use in the ZET. They did however not harmonize their testing methods before starting this project. One one hand this could make interpretation of the data more difficult, but on the other hand it could also provide important information on the most optimal test methods.

Based on the presented data in **Chapter 8**, it can be concluded that bioanalysis to asses ZF exposure should always be included. This would aid in distinction of false negative or false positive results [Cassar, 2017]. Further investigation is needed to define at which internal concentration false results (positive or negative) can be expected. Important to recognize is that internal ZF exposure is directly linked to chemical properties of the test compound which was investigated in **Chapter 9**. We were the first to measure the actual distribution of test compounds within the ZF egg every 24 hours, taking the amount found in the embryo and the chorion as separate parts of the complete egg. We showed that logP value not only highly influences compound uptake in the complete egg, but we also found a striking difference between compound distribution within ZF embryos between low and high logP values. These results are of great importance for data interpretation regarding compound toxicity as Padilla and collegues found an overall pattern of increasing toxicity likelihood with increasing logP [Padilla 2012]. They concluded that both the likelihood that a chemical will be toxic to the developing zebrafish embryo and the degree of toxicity (i.e., AC50) were related to the hydrophobicity of the chemical.

However, logP value alone will not predict the compounds uptake within the ZF as next to logP value, factors like molecular weight, molecular charge, and molecular structure could also highly influence uptake and thus internal exposure within the ZF embryo. Pelka et al recently documented a strong size-dependency of the uptake of substances across the chorion of ZF embryos [Pelka, 2017]. They concluded that the size limit for the uptake of chemicals such as polyethylene glycols into the ZF egg is about 3 kDa for embryos younger than 24 h and might at least partially reach 4 kDa in 48 h embryos.

Therefore, to measure and quantify internal exposure, it is imperative to apply bioanalysis additional to tests done with ZF embryos. This should be done on at least two occasions (i.e. before and after hatching), and it is essential to manually dechorionate the egg before hatching in order to determine the concentration within the embryo as opposed to the complete egg.

In addition to above described hurdles, there is another challenge to overcome. It should be realized that a major difference between developmental toxicity testing in rats and rabbits versus ZF, is the lack of the dam and placenta. This might affect the absorption, distribution, metabolism and excretion of a test compound. It is mportant to realize that the zebrafish embryo by default lives from its yolk only and is therefore not inclined to absorb compounds from external sources, whereas the mammalian embryo lives in a highly vascularized environment because it needs to get its nutrition from the mother, and thus is more likely (dependent on maternal kinetics) to absorb compounds from external sources. The lack of a maternal barrier during ZF development implies direct exposure of the embryo to the parent compound in teratogenicity assays, while mammalian embryos/fetuses are exposed to the parent compound and its metabolites due to drug metabolism by, predominantly, the dam's liver [Verbeuken, 2017]. Hence, ZF embryos depend on their own drug-metabolizing capacity for detoxification and/or bioactivation of xenobiotics. The latter is particularly important for compounds that require bioactivation to exert their teratogenic potential, i.e., so-called proteratogens. A lack of intrinsic biotransformation in the ZF embryo can lead to false negative results in teratogenicity assays as proteratogens will be missed. Recently, Verbeuken et al showed that the capacity of intrinsic biotransformation in ZF embryos appears to be lacking during a major part of organogenesis [Verbeuken, 2017], which is a major shortcoming of this system and should be taken into account when evaluating the study results. Moreover, until now the use of a metabolic activation system (MAS) has been unsuccessful, as the current system is embryotoxic and consequently ZF embryos are only exposed during a very limited developmental window [Pype, 2017].

In order to obtain valid ZET results that are human relevant, I would insist to always include bioanalyses (as described above) and to take metabolism of the test compound into account. In case no human metabolism data exist, this should be evaluated by other means (e.g. in silico models).

10.2 FUTURE PERSPECTIVES IN REGULATORY DART TESTING

From the results in these chapters it can be concluded that, in general, recent applied ethical improvements in regulatory accepted DART testing have focused on reduction and refinement. These are major ethical improvements, and we should continue to expand on this. Moreover, ongoing scientific and technical innovations will provide the means for replacement of animal testing in the future, and thereby reducing animal testing to a minimum and realizing the supreme ambition towards animal-free hazard and risk assessment. The greatest revolution that will define the next era of toxicity testing is the shift from tick-box driven standard animal experiments for regulatory safety assessment towards tailor-made approaches by intelligent and pragmatic testing using computational, molecular, and in vitro tools. This change will reflect the revolution in scientific, technological and regulatory knowledge, and will provide an improved model for human risk assessment.

In the coming years, there are still important improvements to make and challenges to overcome. These can be divided in three parts: 1) refinement of testing methods, 2) reduction of animal use, and 3) replacement of animal testing.

10.2.1. Refinement of testing methods

Bioanalytical method development to align with CMS

Capillary microsampling (CMS) has many advantages as it serves refinement as well as reduction. It can be used in all study types and animal species (i.e. small and large). However, when using these small sample sizes, the bioanalytical method should be developed with this aim and be sufficiently sensitive. In case this was not done at the start of preclinical testing, the companies will usually not redo their work and a huge opportunity to refine and reduce will be missed. It is therefore essential to create awareness of this fact for all people involved [Chapman 2014a, Chapman 2014b].

Refined selection of high dose level by use of TK data

Currently, dose selection for DART studies is guided by the need to demonstrate some maternal toxicity (e.g., adverse clinical signs, deficits in body weight gain and/or food consumption) at the high dose. Although there is a limit dose of 1000 mg/kg body weight/day included in test guidelines for pharmaceuticals, agrochemicals and industrial chemicals, this dose would often represent exposure that is much higher than the anticipated human exposure. Integrated TK data give details on the rate, extent and duration of systemic exposure across doses, species, strains, gender and developmental phases within a toxicology setting. Use of TK can aid to ensure that studies are designed to be relevant for risk assessment in humans, and circumvent using disproportionately high doses that could result in preventable suffering in experimental animals [Brannen 2011]. Moreover, in case of absent toxicity at the limit dose of 1000 mg/kg body weight/day, TK data will provide valuable information on whether the animal model was adequately exposed.

To improve the relevance of these studies for human risk assessment and to reduce preventable animal suffering, it would be of great value to select the high dose level based on TK data and thereby avoiding dose levels that result in nonlinear kinetics (when these are not relevant for risk assessment e.g. with low environmental exposure) [Creton 2012].

It is important to realize that the use of maternal toxicity for dose level assessment in DART studies was selected as a way to enhance the sensitivity of these studies, by increasing the chance for detecting treatment-related abnormalities. However, this approach assumes that findings that appear at high doses are relevant to low doses, which is often not the case. In this line, several years ago in 2001, ILSI/HESI-ACSA technical committee proposed a tiered approach with emphasis on determining systemic dose of parent and/or metabolite(s) in test animals as biological effects are reflective of systemic rather than administered dose [ILSI/HESI 2001]. It was proposed that toxicity studies should ideally be conducted at doses in the linear range of external to internal dose conversion, or slightly above the point of departure from linearity or kinetically-derived maximum dose (KMD) as the systemic dose nonlinearity is a more sensitive parameter occurring much ear-

lier than the maximum-tolerated-dose (MTD) endpoints. Determining systemic dose in studies will also increase the understanding of the mode-of-action (MoA) and reduce the use of animals by avoiding unnecessary additional MoA studies [Saghir 2015].

Also for DART studies, TK evaluations can greatly enrich the value of the generated data and is by default used in safety assessment of pharmaceuticals. Recently, exposure control analyses are also included in DART testing of agrochemicals. For industrial chemicals these analyses are only rarely performed, but currently they are getting more attention as shown by increased recommendations of these evaluations in OECD test guidelines (e.g. OECD 443, EOGRTS) and literature [Marty 2013, Creton 2012].

In the future, when TK will be applied more frequently it will automatically provide details for implementation of the recommended approach for dose level selection based on systemic dose rather than administered dose. This will greatly enhance the ethical use of laboratory animals and will result in more relevant data for human risk assessment. Therefore, it will be important to continue promoting the important role TK can play in chemical risk assessment, with the ultimate goal to ensure incorporation as regulatory requirement. Moreover, the use of CMS would ensure no use of additional animals for this purpose and the least discomfort.

10.2.2. Reduction of animal use

Added value of second species in EFD testing

One topic that was not touched upon in this thesis, is the added value of a second species in developmental toxicity studies. Based on the available data from retrospective analyses mainly for pharmaceuticals, food additives, agrochemicals, and veterinary products, it is clear that for these compounds the use of both species has a higher probability of detecting developmental toxicants than either one alone [Theunissen 2016a, Theunissen 2016b, Janer 2008, Hurtt 2003, Knudsen 2009]. In this respect, there are four important matters that need further investigation:

1) Additional analyses are necessary to evaluate if this conclusion also applies for industrial chemicals,
2) It will be worthwhile to investigate if the two species should be the rat and rabbit, or if one mammalian study can be replaced by an alternative species (e.g. the zebrafish),
3) Instead of performing both developmental studies at once, a tiered approach should be followed, and
4) It would save a lot of animals in case a developmental toxicity study in the rat will only be performed if triggered based on concern from data of a generation/littering study in the rat and developmental study in the rabbit (as proposed for agrochemicals).

Requirement of EOGRTS for agrochemicals

Commission Regulation (EU) No. 283/2013 requests a two-generation study for agrochemicals, however it is also stated that the OECD extended one-generation reproductive toxicity study may

be considered as an alternative approach to the multi-generation study [Commission Regulation (EU) No 283/2013]. Noteworthy is that the original idea for an EOGRTS was proposed by the Agricultural Chemical Safety Assessment (ACSA) Technical Committee of the ILSI Health and Environmental Sciences Institute (HESI) for an improved approach to assess the safety of crop protection chemicals [Cooper, 2006], which is now implemented as final guideline for industrial chemicals but not yet for agrochemicals. However Commission Regulation (EU) No. 283/2013 also mentions that potential neurotoxic, immunotoxic effects and effects potentially related to changes in the hormonal system shall be carefully addressed and reported [Commission Regulation (EU) No 283/2013]. Especially these items were added to the EOGRTS design. As long as the Commission Regulation is not updated to explicitly request an EOGRTS instead of two-generation study (and since the two-generation study is still requested by the US EPA), it is anticipated that the majority of agrochemicals will be tested in two-generation studies as this would save costs substantially for the agrochemical companies, however in respect of the 3Rs this would be the least favorable.

10.2.3. Replacement of animal testing

Integrated Approaches to Testing and Assessment (IATA)

One concept that is getting considerable attention for toxicity testing in the future and would cause the biggest change, is the shift from a standard list of animal experiments for regulatory safety assessment towards tailor-made approaches by intelligent and pragmatic testing using computational, molecular, and in vitro tools [Krewski 2010, Adeleye 2015]. Recently implemented changes have focused on reduction and refinement. The current concept aims for replacement of animal testing which will cause the biggest revolution in toxicology testing.

The need for new approaches was discussed in “Toxicity Testing in the 21st Century: A Vision and a Strategy” by the National Research Council (NRC) in 2007 [NRC 2007]. In this report, the use of in vitro systems and pathway analysis was advocated, which would allow more efficient and high-throughput screening methods. The report points out that we should take advantage of the revolutions in biology and biotechnology, including toxicogenomics, bioinformatics, systems biology, epigenetics and computational toxicology. Since this publication, extensive ambition has grown worldwide, also in regulatory bodies, to have more relevant testing strategies which also support the 3R principles. This concept was created for toxicity testing in general, not specific for DART.

However, especially for DART, it will not be possible to interchange complex in vivo tests by relatively simple in vitro tests, as DART testing is concerned with all phases of the reproductive cycle, starting with female and male fertility, prenatal and postnatal development, and culminating in late manifestations that can only be detected in the next generation. The reproductive cycle constitutes an extensive list of mechanisms on the molecular, cellular, tissue and organism level, with different windows of sensitivity over time. Due to its complexity, the whole system cannot be modelled in one in vitro test; the different components of the reproductive cycle will have to be studied individually and then integrated into testing strategies.

One of the initiatives following the NRC report is Toxicology in the 21st Century (Tox21). This is a federal collaboration among EPA, NIH, including National Center for Advancing Translational Sciences and the National Toxicology Program at the National Institute of Environmental Health Sciences, and the Food and Drug Administration. Tox21 researchers aim to develop better toxicity assessment methods to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to negative health effects [Merrick 2015, Attene-Ramos 2013, Tice 2013]. One of EPA's contributions to Tox21 are the chemical screening results from the Toxicity Forecaster (i.e. ToxCast). ToxCast has data on more than 1,800 chemicals from a broad range of sources and these are screened in over 700 high-throughput assays that cover a range of high-level cell responses and approximately 300 signaling pathways [Richard 2016, Judson 2014]. As all ToxCast chemical data is publicly available, it is extremely valuable for the development and validation of novel predictive models. The goal of Tox21 is to develop more efficient and less time-consuming approaches to predict how chemicals may affect human health. Initially, the main focus of Tox21 is to help prioritize chemicals for more extensive testing using traditional methods. Tox21's ultimate aim is to develop strategies for agencies to use in regulating chemicals and reducing the current reliance on animal testing for toxicological evaluations.

To use this type of information in an effective way, the conceptual framework of Adverse Outcome Pathway (AOP) is proposed. An AOP portrays existing knowledge concerning the linkage between two anchor points: the Molecular Initiating Event (MIE) and the Adverse Outcome (AO), at a level of biological organization relevant to risk assessment [Knapen 2015, Villeneuve 2014, Tollefsen 2014]. The AOP concept can be applied as a framework to develop Integrated Approaches to Testing and Assessment (IATA). Such a framework will usually consist of three main elements: the AOP itself, non-animal (alternative) test methods and in silico approaches targeting key components of the AOP, and their associated prediction models for a particular regulatory context.

Important scientific and technical innovations that are essential for this concept are in vitro models, omics technologies and computational or in silico technologies.

In vitro models will provide the main biological systems for toxicity testing instead of the currently used whole-animal models. For developmental toxicity testing, the rodent whole embryo culture, zebrafish embryo assay, and embryonic stem cell assay are the most popular [Zhang 2016].

Omics technologies have the goal of analyzing the components of a living system as a whole. The omics field falls into various subdivisions, depending on the cellular component being measured, including the genome, transcriptome, proteome, and metabolome. The breadth and sensitivity of these technologies provide great opportunities to identify molecular perturbations leading to toxicity, and therefore omics tools will play an important role in establishing the proposed toxicity pathway-based risks assessment [Merrick 2015, van Vliet 2011].

Computational or in silico biology uses computer techniques and mathematical modeling to understand biologic processes. It is a powerful tool to cope with the increasing quantity and quality of biologic information (like toxicogenomics and metabolomics), but also for modelling purposes (e.g. for kinetics). Application of in silico models to predict developmental toxicity has demonstrated limited success particularly when employed as a single source of information [Marzo 2016], and therefore it should be used together with a combination of multiple experiments as is done for ToxCast.

As the AOP and IATA frameworks are increasingly being used for 'simple toxicity endpoints' (like skin sensitization), it will become an interesting challenge to develop this further for more complex endpoints like DART. A good starting point is the list of 17 developmentally important molecular signaling pathways which are used repeatedly throughout organogenesis that was published by the Committee on Developmental Toxicology, Board on Environmental Studies and Toxicology, National Research Council in 2000 [NRC 2000]. Important to know is that these 17 signaling pathways are highly conserved across numerous animal phyla (e.g., nematode worms to arthropods to chordates). The molecular targets and responses within cells also are conserved across phyla, including specific gene expression, cell migration, and cell proliferation. The differences in the development of various organisms mostly reflect differences in the particular times, places, and combinations of use of the conserved pathways and responses. Those findings give new validity to the use of model organisms to learn more about basic development in mammals, to provide mechanistic clues about human variability, and to analyze and assess the risks of potential developmental toxicants [NRC 2000].

It is important to select a group of complementary tests (in a tiered or battery approach), which contains all applicable mechanisms of action [Sogorb 2014, Piersma 2014]. The combined outcome for a particular substance in the complete battery will provide vital details on mechanisms of action of the test substance which is important for extrapolation to human relevance (because it considers toxic mechanisms rather than endpoints affected). The concept of an alternative in vitro test battery was recently successfully tested within the concept of grouping and read across to distinguish in vivo non- or weak developmental toxicants from potent developmental toxicants within groups of structural analogs [Kroese 2015]. This study clearly illustrates the added value of combining assays that contain complementary biological processes and mechanisms, increasing the predictive value of a screening battery over individual in vitro assays. However, given the large number of mechanistic aspects related to the reproductive cycle, an in vitro test battery a priori cannot be expected to be able to detect all reproductive and developmental toxicants. This is just a first step and it must be stressed that embryonic development includes both key molecular signaling mechanisms and multifaceted organization of biological systems. In addition, homeostasis and repair capability impact the final developmental outcome and should be considered in the formation of the alternative testing plans.

10.3. CONCLUSION

The main conclusions from this thesis are:

- Refined blood sampling techniques should be incorporated in DART studies, leading to a reduction in animal use.
- The recent changes in OECD guidelines for reproduction studies impact on refinement and reduction. Additional information is gathered which improves the value of these studies, and less animals are used in the EOGRTS as no second generation is included by default. However, the number of animals is slightly increased in repro screening studies.
- The zebrafish embryotoxicity test (ZET) is a first step towards replacement in future DART testing, though optimization and harmonization is needed. Moreover, it is important to select a group of complementary tests instead of searching for one-to-one replacements.
- Practical and scientific knowledge on the performance of regulatory DART studies should be published to improve and harmonize these methods worldwide with special focus on refinement, reduction and replacement.

Overall, already many animal welfare improvements have been made for regulatory DART testing, but these all relate to reduction and refinement. However, the greatest revolution will comprise the replacement of animal testing which should be our ultimate goal. With this thesis we would like to engage scientists in all sectors - government, public interest, industry, university, and consulting laboratories - who design and conduct toxicity tests and who use study data to assess human risk. As several necessary tools are currently available or in an advanced stage of development, but still a lot of work needs to be done, the essential link to accomplish this paradigm shift and to raise to the next level is a dedication to revolutionize by the scientists. The potential advantages are obvious. It is now up to us to make this change.

Appendix 1

Nederlandse samenvatting

NEDERLANDSE SAMENVATTING

Mens, dier en milieu worden, bewust of onbewust, blootgesteld aan verscheidene chemicaliën waarvan de grootste groep bestaat uit industriële chemicaliën, agro chemicaliën en medicijnen. Om de veiligheid voor mens, dier en milieu te waarborgen, worden deze chemicaliën getest op toxiciteit (giftigheid). Veel van deze testen maken gebruik van proefdieren. Er zijn verschillende toxiciteits-testen, en deze moeten uitgevoerd worden afhankelijk van de type stof en de te verwachte blootstelling. Het (productie en/of import) volume van industriële chemicaliën bepaalt welke testen er uitgevoerd moeten worden; hoe hoger het volume hoe uitgebreider de testen zijn. Voor agro chemicaliën moet er altijd een vaste reeks testen worden uitgevoerd. En voor medicijnen is er in principe ook een vaste reeks met testen, maar wordt de timing van de testen uitgevoerd ten opzichte van de fase van de klinische studies (in mensen). Daarbij kan het zijn dat de doelgroep van het medicijn ervoor zorgt dat er meer of minder testen uitgevoerd moeten worden (bv meer testen wanneer het medicijn voor kinderen is).

Voor alle drie type chemicaliën is het nodig om ze te testen op 1) veiligheid op de voortplanting (reproductie) en 2) effect op de ontwikkeling van nakomelingen, behalve bij lage productie/import volumes van industriële chemicaliën. De reproductie-testen worden uitgevoerd met ratten en de ontwikkelingstesten worden uitgevoerd met ratten en konijnen. Voor deze testen worden erg veel proefdieren gebruikt doordat er relatief veel dieren per groep nodig zijn om een effect op reproductie aan te tonen. Daarbij tellen alle nakomelingen in beide type studies ook mee aan het proefdiergebruik. Hierdoor gebruiken dit type toxiciteits-testen de grootste aantallen proefdieren in het complete pakket voor veiligheidsbepaling.

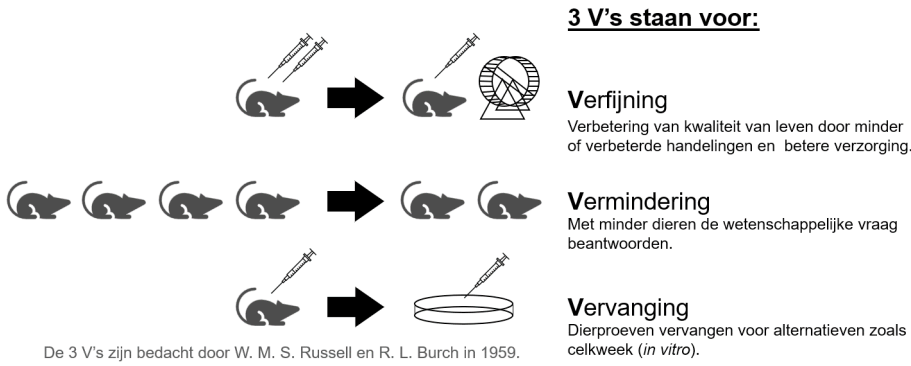
De laatste tijd wordt er meer getwijfeld aan de humane relevantie van deze dierstudies. Mensen zijn immers geen ratten of konijnen.

Er zijn verschillende voorbeelden die laten zien dat een stof ernstige effecten in de rat laat zien, die niet te zien zijn in de mens. Bijvoorbeeld, aspirine veroorzaakt aangeboren afwijkingen in ratten, maar niet in mensen. In de huidige manier van testen wordt er niet gekeken naar de manier waarop bepaalde effecten in proefdieren zijn ontstaan, waardoor je niet weet of de bevinding relevant voor de mens is. Veel winst zal dus te halen zijn door beter te kijken naar hoe bepaalde effecten ontstaan en hoe deze mechanismen hetzelfde zijn in de mens.

Tevens is er een groot verschil in de hoeveelheid blootstelling van de proefdieren versus de mens. De in utero ontwikkeling van een rat duurt 21 dagen en van een mens 9 maanden. Per dag zijn er dus veel meer ontwikkelingsprocessen bezig in de rat wanneer je dit vergelijkt met de mens. Hierdoor kan een dagelijkse hoge dosering op 1 moment een veel ernstiger impact hebben op de ontwikkeling van de rat ten opzichte van de mens.

Daarbij zijn de testen nog niet compleet en ontbraken tot voor kort bepaalde metingen die erg belangrijk zijn voor de mens. Deze zijn recent toegevoegd en bestaan uit testen voor hormoon verstoringen, effecten op gedrags- en hersen ontwikkeling, en op ontwikkeling van het immuunsysteem.

Met het oog op de 3Vs: verfijning, vermindering en vervanging van proefdieren (zie figuur), de grote aantallen proefdieren die gebruikt worden in toxiciteits-testen voor reproductie en ontwikkeling, en de onzekerheden in de humane relevantie van deze testen, valt er veel winst te behalen door dit type studies te optimaliseren.



Figuur. Grafische weergave van de 3Vs: verfijning, vermindering en vervanging van proefdieren.

Het doel van het onderzoek beschreven in dit proefschrift is om te evalueren of en waar het mogelijk is om de 3Vs te implementeren in toxiciteits-testen voor reproductie en ontwikkeling, zonder de integriteit en validiteit van de studie opzet te verliezen. De focus is op het werk in de praktijk waarbij ik collega's wereldwijd enthousiast wil krijgen zodat zij deze verbeteringen ook gaan toepassen.

Het proefschrift bestaat uit 3 delen. Het eerste deel (hoofdstukken 2 en 3) beschrijft de winst die valt te halen door verfijning van de methode voor bloed afname. Hierdoor ondervindt het proefdier minder ongerief en kunnen er zelfs substantieel minder ratten worden ingezet. Het tweede deel (hoofdstukken 4, 5, 6 en 7) beschrijft de impact op verfijning en vermindering in recent aangepaste richtlijnstudies voor industriële chemicaliën. En het derde deel (hoofdstukken 8 en 9) beschrijft de mogelijkheid voor vervanging van huidige testen voor ontwikkelingstoxicologie door de zebravis embryotoxiciteits test.

In **hoofdstuk 2** van dit proefschrift wordt een revolutionaire manier van bloedafname beschreven, namelijk capillaire microsampling (CMS). Het volume van de bloedmonsters kan hierdoor van 0.3 ml naar 0.03 ml worden teruggebracht waardoor het ongerief voor de proefdieren veel minder is.

Wij hebben laten zien dat door deze methode in te voeren bij studies met juveniele ratten, het mogelijk is om het aantal dieren dat bestemd is voor bloedafname (satellietdieren) te verminderen met 75%. **Hoofdstuk 3** laat zien dat in de gevallen wanneer CMS niet gebruikt kan worden, bloedafname via de jugularis ader in ontwikkelingstoxiciteits-testen in de rat ervoor kan zorgen dat er geen satellietdieren nodig zijn. Dit geeft een proefdier vermindering van 20%.

Hoofdstuk 4 van dit proefschrift beschrijft de effectiviteit van reproductie/ontwikkeling screening studies (volgens OECD 421 en 422 richtlijnen uit 1995 en 1996) met een retrospectieve evaluatie van 134 studies. Hieruit hebben we geconcludeerd dat deze type studies erg waardevol zijn doordat ze unieke data genereren met een relatief laag aantal proefdieren, maar dat een paar simpele toevoegingen de risicobeoordeling verder kan verbeteren.

Een aantal van de toevoegingen beschreven in Hoofdstuk 4 zijn geïmplementeerd als verbetering van de OECD 421 en 422 richtlijnen in 2015. Details hiervan zijn beschreven in **Hoofdstuk 5**. Dit hoofdstuk geeft praktische details over de uitvoering van deze extra metingen, om zo onze ervaringen te delen zodat andere laboratoria, studie monitoren en registratie autoriteiten hiervan gebruik kunnen maken. Hoofdstuk 5 vermeld ook een aantal kritische punten met betrekking tot deze richtlijn aanpassing en het effect op de 3Vs.

In **Hoofdstukken 6 en 7** beschrijven we onze ervaring met de implementatie van OECD 443, de uitgebreide 1-generatie studie. Het belangrijkste kenmerk van dit type studie is dat gebaseerd op bepaalde criteria, geselecteerde nakomelingen worden toegewezen aan verschillende cohorten. Deze cohorten worden gebruikt om mogelijke effecten te onderzoeken op ontwikkeling, integriteit en functie van de geslachtsorganen, neuropathologie, gedrag en/of functie van het immuunsysteem. De criteria zorgen voor een op maat gemaakte studie, rechtstreeks gebaseerd op data, terwijl het proefdiergebruik geminimaliseerd wordt. Beide hoofdstukken beschrijven belangrijke kenmerken van deze uitdagende studie, waarbij Hoofdstuk 6 zich richt op praktische aspecten voor uitvoer van de uitgebreide 1-generatie studie, en Hoofdstuk 7 sturing geeft over wetenschappelijk-gebaseerde criteria voor de uitgebreide evaluaties (neurotoxicologie, immunotoxicologie en dekking voor de tweede generatie).

De embryotoxiciteits test in zebravissen is ontwikkeld als alternatief voor ontwikkelingstesten in ratten en konijnen. Op dit moment bestaan er verschillende manieren om de zebravis embryotoxiciteits test uit te voeren. Om harmonisatie wereldwijd in gang te zetten, stellen wij in **Hoofdstuk 8** de meest optimale studie opzet voor. Deze opzet is gebaseerd op artikelen en experimenteel bewijs wat betreft temperatuur, test kamer, blootstellingsperiode, aanwezigheid van het chorion (=membraan om het ei), gebruik van oplosmiddelen, blootstellingsmethode, concentraties, en teratogene classificatie (=classificatie voor ernstige aangeboren afwijking). Tevens bevat dit hoofdstuk onze methode voor morfologische gradering om harmonisatie nog verder te helpen. **Hoofdstuk 9** beschrijft de essentiële rol van het bepalen van de interne blootstelling op de voorspellende uitkomst van de zebravis embryotoxiciteits test. Het effect van lipofiliciteit (vetoplosbaarheid), geb-

ruik van oplosmiddelen en de aanwezigheid van het chorion op de distributie van de teststof in de zebravis embryo zijn hierin gerapporteerd.

Van de resultaten in dit proefschrift kan geconcludeerd worden dat, in het algemeen, recent toegepaste ethische verbeteringen in ontwikkelings- en reproductietoxicologie testen volgens de regelgeving gefocust zijn op proefdier verminderen en verfijning. Dit zijn belangrijke ethische verbeteringen en het is belangrijk om dit te blijven doen. Daarbij zullen nieuwe wetenschappelijke en technische innovaties de middelen leveren om proefdiervervanging te implementeren. Hierdoor zal in de toekomst het aantal proefdier experimenten geminimaliseerd worden, met de ultieme ambitie om uiteindelijk de gevaar- en risicoschatting volledig zonder proefdier experimenten uit te voeren.

Appendix 2

References

REFERENCES

1. 6341 ISO. 1996. Water quality - Determination of the inhibition of the mobility of Daphnia magna Straus (Cladocera, Crustacea) - Acute toxicity test.
2. Acciani M, Kopp C, Palmer C, Davis T, Picut C (2015) Effects of immersion fixation on the post-mortem rat brain. Poster, Society of Toxicology Annual Meeting, San Diego, CA.
3. Adeleye Y, Andersen M, Clewell R, Davies M, Dent M, Edwards S, Fowler P, Malcomber S, Nicol B, Scott A, Scott S, Sun B, Westmoreland C, White A, Zhang Q, Carmichael PL. Implementing Toxicity Testing in the 21st Century (TT21C): Making safety decisions using toxicity pathways, and progress in a prototype risk assessment. Toxicology. Jun 5;332:102-11. doi: 10.1016/j.tox.2014.02.007 (2015).
4. Adler S, Pellizzer C, Hareng L, Hartung T, Bremer S. First steps in establishing a developmental toxicity test method based on human embryonic stem cells. Toxicol In Vitro. 2008 Feb;22(1):200-11. Epub 2007 Sep 4.
5. Ai, W., Huo, Y., Liu, X., Liu, F., Zhou, X., Miao, Y., Jiang, H., Zhang, L., Shen, L., Piao, J., Li, B., 2014. Relative sensitivities of TDAR, cytokine production, and immunophenotyping assays in immunotoxicity assessment. Toxicol. Res. 3, 465–473.
6. Ai, W., Li, H., Song, N., Li, L., Chen, H., 2013. Optimal method to stimulate cytokine production and its use in immunotoxicity assessment. Int. J. Environ. Res. Public Health 10, 3834–3842.
7. Al Jishi T, Sergi C. Current perspective of diethylstilbestrol (DES) exposure in mothers and offspring. Reprod Toxicol. 2017 Apr 28;71:71-77. doi: 10.1016/j.reprotox.2017.04.009.
8. Ali S, van Mil HGJ, Richardson MK. Large scale assessment of the zebrafish embryo as a possible predictive model in toxicity testing. Plos One. 2011. 6(6):1-11.
9. Alwan S, Chambers CD. Identifying Human Teratogens: An Update. J Pediatr Genet. 2015 Jun;4(2):39-41. doi: 10.1055/s-0035-1556745.
10. Anon, 2006. The Innovative Medicines Initiative (IMI) Strategic Research Agenda. European Federation of Pharmaceutical Industries and Associations.
11. Arnold M, Langhans W. Effects of anesthesia and blood sampling techniques on plasma metabolites and corticosterone in the rat. Physiol Behav. 2010 Apr 19;99(5):592-8. doi: 10.1016/j.physbeh.2010.01.021. Epub 2010 Feb 10.
12. Arnot JA, Arnot MI, Mackay D, Couillard Y, MacDonald D, Bonnell M, Doyle P. 2010. Molecular size cut-off criteria for screening bioaccumulation potential: fact or fiction? Integr Environ Assess Manag 6(2): 210-224.
13. Attene-Ramos MS, Miller N, Huang R, Michael S, Itkin M, Kavlock RJ, Austin CP, Shinn P, Simeonov A, Tice RR, Xia M. The Tox21 robotic platform for the assessment of environmental chemicals--from vision to reality. Drug Discov Today. Aug;18(15-16):716-23. doi: 10.1016/j.drudis.2013.05.015 (2013).
14. Aulmann W. Assessment of reproductive toxicity under REACH. Regul Toxicol Pharm 2012;63:286–90.
15. Ball JS, Stedman DB, Hillegass JM, Zhang CX, Panzica-Kelly J, Coburn A, Enright BP, Tornesi B, Amouzadeh HR, Hetheridge M, Gustafson AL, Augustine-Rauch KA. Fishing for teratogens: a consortium effort for a harmonized zebrafish developmental toxicology assay. Toxicological sciences. 2014. 139(1):210-9.
16. Barten, M.J., Rahmel, A., Bocsi, J., Boldt, A., Garbade, J., Dhein, S., Mohr, F.W., Gummert, J.F., 2006. Cytokine analysis to predict immunosuppression. Cytometry A 69, 155–157.
17. Beasly A, Elrod-Erickson M, Otter RR. Consistency of morphological endpoints used to assess developmental timing in zebrafish (Danio rerio) across a temperature gradient. Reproductive toxicology. 2012. 34:561-7.
18. Beekhuijzen M, Barentsen H, Marsden E, Zmarowski A, Aujoulat M, Picut C, Sloter E (2016) Implementing the Extended One-Generation Reproductive Toxicity Study (EOGRTS). Important points to consider. Crit Rev Toxicol Apr;46(4):332-47. doi: 10.3109/10408444.2015.1137863
19. Beekhuijzen M, de Koning C, Flores-Guillén ME, de Vries-Buitenweg S, Tobor-Kaplon M, van de Waart B, Emmen H. 2015. From cutting edge to guideline: A first step in harmonization of the zebrafish embryotoxicity test (ZET) by describing the most optimal test conditions and morphology scoring system. Reprod Toxicol S0890-6238(15):00173-2.
20. Beekhuijzen M, de Raaf MA, Zmarowski A, van Otterdijk F, Peter B, Emmen H (2014) The underestimated value of OECD 421 and 422 repro screening studies: Putting it in the right perspective. Reprod Toxicol 48:81-87. doi: 10.1016/j.reprotox.2014.04.003

21. Beekhuijzen, M., Zmarowski, A., Emmen, H., Frieling, W., 2009. To mate or not to mate: a retrospective analysis of two-generation studies for evaluation of criteria to trigger additional mating in the extended one-generation design. Reprod. Toxicol. 28, 203–208.
22. Berghmans S, Butler P, Goldsmith P, Waldron G, Gardner I, Golder Z, Richards FM, Kimber G, Roach A, Alderton W, Fleming A. 2008. Zebrafish based assays for the assessment of cardiac, visual and gut function - potential safety screens for early drug discovery. Journal of Pharmacological and Toxicological Methods 58: 59-68.
23. Blazak, W.F., 1989. Significance of cellular end points in assessment of male reproductive toxicity, in: Working P.K. (Ed.), Toxicology of the Male and Female Reproductive Systems. Hemisphere Publishing, New York, pp. 157–172.
24. Bolon B, Garman R, Jensen K, Krinke G, Stuart B, Ad Hoc Working Group of the STP Scientific and Regulatory Policy Committee (2006) A 'best practices' approach to neuropathologic assessment in developmental neurotoxicity testing for today. Toxicol Pathol 34(3):296-313. doi: 10.1080/01926230600713269
25. Boverhof, D.R., Ladics, G., Luebke, B., Botham, J., Corsini, E., Evans, E., Germolec, D., Holsapple, M., Loveless, S.E., Lu, H., van der Laan, J.W., White, K.L. Jr, Yang, Y., 2014. Approaches and considerations for the assessment of immunotoxicity for environmental chemicals: a workshop summary. Regul. Toxicol. Pharmacol. 68:96–107.
26. Brannen KC, Fenton SE, Hansen DK, Harrouk W, Kim JH, Shuey D. Developmental toxicology – new directions workshop: refining testing strategies and study designs. Birth Defects Res B Dev Reprod Toxicol. 2011 October; 92(5): 404–412. doi:10.1002/bdrb.20326.
27. Brannen KC, Panzica-Kelly JM, Danberry TL, Augustine-Rauch KA. Development of a zebrafish embryo teratogenicity assay and quantitative prediction model. Birth defects research (Part B). 2010. 89:66-77.
28. Braunbeck T, Boettcher M, Hollert H, Kosmehl T, Lammer E, Leist E, Rudolf M, Seitz N. Towards an alternative for the acute fish LC(50) test in chemical assessment: the fish embryo toxicity test goes multispecies—an update. ALTEX. 2005. 22:87–102.
29. Brox S, Ritter AP, Küster E, Reemtsma T. 2014. Influence of the perivitelline space on the quantification of internal concentrations of chemicals in eggs ofzebrafish embryos (Danio rerio). Aquatic Toxicology 157: 134-140.
30. Busquet F, Nagel R, von Landenberg F, Mueller SO, Huebler N, Broschard TH. Development of a new screening assay to identify proteratogenic substances using zebrafish Danio rerio embryo combined with an exogenous mammalian metabolic activation system (mDarT). Toxicol. Sci. 2008. 104:177–88.
31. Busquet F, Strecker R, Rawlings JM, Belanger SE, Braunbeck T, Carr GJ, Cenijn P, Fochtman P, Gourmelon A, Hubler N, Kleensang A, Knobel M, Kussatz C, Legler J, Lillicrap A, Martinez-Jeronimo F, Polleichtner C, Rzodeczko H, Salinas E, Schneider KE, Scholz S, van den Brandhof EJ, van der Ven LTM, Walter-Rohde S, Weigt S, Witters H, Halder M. OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. Regulatory Toxicology and Pharmacology. 2014. 69:496–511.
32. Carney EW, Zablotny CL, Marty MS, Crissman JW, Anderson P, Woolhiser M, Holsapple M (2004) The effects of feed restriction during in utero and postnatal development in rats. Toxicol Sci 82:237–249. doi: 10.1093/toxsci/kfh249
33. Caron A, Lelong C, Bartels T, Dorchies O, Gury T, Chalier C, Benning V. 2015. Clinical and anatomic pathology effects of serial blood sampling in rat toxicology studies, using conventional or microsampling methods. Regul Toxicol Pharmacol 72(3):429–439.
34. Cassar S, Breidenbach L, Olson A, Huang X, Britton H, Woody C, Sancheti P, Stolarik D, Wicke K, Hempel K, LeRoy B. Measuring drug absorption improves interpretation of behavioral responses in a larval zebrafish locomotor assay for predicting seizure liability. J Pharmacol Toxicol Methods. 2017 Jul 13;88(Pt 1):56-63. doi: 10.1016/j.vascn.2017.07.002. [Epub ahead of print].
35. Chang J, Wang M, Gui W, Zhao Y, Yu L, Zhu G. Changes in thyroid hormone levels during zebrafish development. Zoological Science. 2012. Volume 29(3):181-4.
36. Chapin R, Augustine-Rauch K, Beyer B, Daston G, Finnell R, Flynn T, Hunter S, Mirkes P, O'Shea KS, Piersma A, Sandler D, Vanparys P, Van Maele-Fabry G. State of the art in developmental toxicity screening methods and a way forward: a meeting report addressing embryonic stem cells, whole embryo culture, and zebrafish. Birth defects research (Part B). 2008. Aug;83(4):446-56.

37. Chapin, R.E., Gulati, D.K., Barnes, L.H., Teague, J.L., 1993. The effects of feed restriction on reproductive function in Sprague–Dawley rats. Fundam. Appl. Toxicol. 20, 23–29.
38. Chapman K, Burnett J, Corvaro M, Mitchell D, Robinson S, Sangster T, Sparrow S, Spooner N, Wilson A. 2014a. Reducing pre-clinical blood volumes for toxicokinetics: toxicologists, pathologists and bioanalysts unite. Bioanalysis 6(22):2965–2968.
39. Chapman K, Chivers S, Gliddon D, Mitchell D, Robinson S, Sangster T, Sparrow S, Spooner N, Wilson A. 2014b. Overcoming the barriers to the uptake of nonclinical microsampling in regulatory safety studies. Drug Discov Today 19 (5):528–532.
40. Cheng J, Flahaut E, Cheng SH. 2007. Effect of carbon nanotubes on developing zebrafish (Danio rerio) embryos. Environ Toxicol Chem 26(4): 706-716.
41. Clouet E, Kerdine-Römer S, Ferret PJ. Comparison and validation of an in vitro skin sensitization strategy using a data set of 33 chemical references. Toxicol In Vitro. 2017 May 21. pii: S0887-2333(17)30130-3. doi: 10.1016/j.tiv.2017.05.014. [Epub ahead of print]
42. Clubb, S.K. and Jardine, L. (2015). Clinical chemistry sampling and assessment in juvenile rats. Reduction in and/or elimination of the need for additional subsets of pups. SOT, March, 2015, San Diego, USA (Charles River group, Trenant, UK)
43. Commission Regulation (EU) No 283/2013 of 1 March 2013, setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. Official Journal of the European Union L 93, 1-84, 2013.
44. Commission, E., 2009. Regulation (EC) No 1223/2009 of the European Parliment and of the Council of 30 November 2009 on cosmetics products, Official Journal of the European Union.
45. Connell DW, Hawker DW. 1988. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. Ecotoxicol Environ Saf 16: 242-257.
46. Cooper RL, Kavlock RJ. Endocrine disruptors and reproductive development: a weight-of-evidence overview. J Endocrinol. 1997 Feb;152(2):159-66.
47. Cooper RL, Lamb JC, Barlow SM, Bentley K, Brady AM, Doerrer NG, Eisenbrandt DL, Fenner-Crisp PA, Hines RN, Irvine LF, Kimmel CA, Koeter H, Li AA, Makris SL, Sheets LP, Speijers G, Whitby KE (2006) A tiered approach to life stages testing for agricultural chemical safety assessment. Crit Rev Toxicol 36(January (1)):69–98.
48. Creton R. The calcium pump of the endoplasmic reticulum plays a role in midline signaling during early zebrafish development. Dev. Brain Res. 2004. 151:33-41.
49. Creton S, Saghir SA, Bartels MJ, Billington R, Bus JS, Davies W, Dent MP, Hawksworth GM, Parry S, Travis KZ. Use of toxicokinetics to support chemical evaluation: informing high dose selection and study interpretation. Regul. Toxicol. Pharmacol. 62, 241–247 (2012).
50. Crofton KM, Foss JA, Hass U, Jensen KF, Levin ED, Parker SP (2008) Undertaking positive control studies as part of developmental neurotoxicity testing: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. Neurotoxicol Teratol 30(4):266-87
51. Crofton KM. Thyroid disrupting chemicals: mechanisms and mixtures. Int J Androl. 2008 Apr;31(2):209-23. doi: 10.1111/j.1365-2605.2007.00857.x. Epub 2008 Jan 22.
52. Cunningham, M., Gilkeson, G., 2011. Estrogen receptors in immunity and autoimmunity. Clin. Rev. Allergy Immunol. 40, 66–73.
53. Dang, Z.C., Rorije, E., Esch, T.H., Muller, A., Hakkert, B.C., Piersma, A.H., 2009. Retrospective analysis of relative parameter sensitivity in multi-generation reproductive toxicity studies. Reprod. Toxicol. 28, 196–202.
54. Daston GP, Beyer BK, Carney EW, Chapin RE, Friedman JM, Piersma AH, Rogers JM, Scialli AR. Exposure-based validation list for developmental toxicity screening assays. Birth Defects Res B Dev Reprod Toxicol. 2014 Dec;101(6):423-8. doi: 10.1002/bdrb.21132. Epub 2014 Dec 4.
55. David RM, Jones HS, Panter GH, Winter MJ, Hutchinson TH, Kevin Chipman J. 2012. Interference with xenobiotic metabolic activity by the commonly used vehicle solvents dimethylsulfoxide and methanol in zebrafish (Danio rerio) larvae but not daphnia magna. Chemosphere 88: 912-917.
56. Dawson, A.B. A note on the staining of skeletons of cleared specimens with Alizarin Red S. Stain Technology.1, 123-124 (1926).

57. Delvecchio C, Tiefenbach J, Krause HM. The Zebrafish: A powerful platform for in vivo, HTS drug discovery. ASSAY and Drug Development Technologies. 2011;9(4):354-61.
58. Diehl KH, Hull R, Morton D, Pfister R, Rabemampianina Y, Smith D, Vidal JM, van de Vorstenbosch C. European Federation of Pharmaceutical Industries A, European Centre for the Validation of Alternative M. 2001. A good practice guide to the administration of substances and removal of blood, including routes and volumes. J Appl Toxicol 21(1): 15–23.
59. Dietert, R.R., Holsapple, M.P., 2007. Methodologies for developmental immunotoxicity (DIT) testing. Methods 41, 123–131.
60. Dillen, L., Loomans, T., Van de Perre, G., Versweyveld, D., Wuyts, K, and de Zwart, L. (2014). Blood microsampling using capillaries for drug-exposure determination in early preclinical studies: beneficial strategy to reduce blood sample volumes. Bioanalysis 6, 293-306. doi: 10.4155/bio.13.286
61. Doshna C, Benbow J, DePasquale M, Okerberg C, Turnquist S, Stedman D, Chapin B, Sivaraman L, Waldron G, Navetta K, Brady J, Banker M, Casimiro-Garcia A, Hill A, Jones M, Ball J, Aleo M. Multi-phase analysis of uptake and toxicity in zebrafish: relationship to compound physical-chemical properties. 48th annual meeting of the society of toxicology. Toxicological sciences. 2009;108(S-1):78. #377.
62. Ducharme NA, Peterson LE, Benfenatti E. Meta-analysis of toxicity and teratogenicity of 133 chemicals from zebrafish developmental toxicity studies. Reproductive toxicology. 2013. 41:98-108.
63. Duffell SJ, Soames AR, Gunby S (2000) Morphometric analysis of the developing rat brain. Toxicol Pathol 28:157-163
64. Dunn, O.J. Multiple comparisons using rank sums. Technometrics, 6(3) 241-252 (1964).
65. Dunnett, C.W. A Multiple Comparison Procedure for Comparing Several Treatments with a Control, J. Amer. Stat. Assoc. 50, 1096-1121 (1955).
66. EC 1272/2008. CLP Regulation (1272/2008): EC, Regulation (EC) 1272/2008 of the European Parliament and of the Council on e of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) 1907/2006, 16 December 2008.
67. EC 1907/2006. Regulation (EC) No 1907/2006: EC, Regulation (EC) 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) 793/93 and Commission Regulation (EC) 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, 18 December 2006.
68. ECETOC, 2002. Guidance on Evaluation of Reproductive Toxicity Data. Monograph No 31, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels.
69. ECETOC, 2008. Triggering and Waiving Criteria for the Extended One-Generation Reproduction Toxicity Study. Document No 45, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels.
70. ECETOC, 2008. Workshop on Triggering and Waiving Criteria for the Extended One-Generation Reproduction Toxicity Study, 14-15 April 2008, Barza d'Ispra. Workshop Report No 12, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels.
71. ECHA, Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7a: Endpoint specific guidance, Version 4.1, October 2015
72. ECHA. Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7a. Endpoint Specific Guidance, Version 5.0; December 2016.
73. Ellis-Hutchings RG, Rasoulpour RJ, Terry C, Carney EW, Billington R. Human relevance framework evaluation of a novel rat developmental toxicity mode of action induced by sulfoxaflor. Crit Rev Toxicol. 2014 May;44 Suppl 2:45-62. doi: 10.3109/10408444.2014.910752.
74. EMEA (2013). European Medicines Agency (EMEA), Committee for Human Medicinal Products (CHMP). Guideline on the need for nonclinical testing in juvenile animals on human pharmaceuticals for pediatric indications. 2013.
75. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.3100. 90-Day oral toxicity in rodents. August 1998
76. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.3550, Reproduction/developmental toxicity screening test, July 2000.
77. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.3650, Combined repeated dose toxicity study with the reproduction/developmental toxicity screening test, July 2000.

78. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.3700, Prenatal developmental toxicity study, August 1998.
79. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.3800, Reproduction and fertility effects, August 1998.
80. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.6300. Developmental neurotoxicity study. August 1998
81. EPA (US Environmental Protection Agency). Health Effects Test Guidelines. OPPTS 870.7800. Immunotoxicity. August 1998
82. EPA Source: http://www.epa.gov/reviewing-new-chemicals-under-toxic-substances-control-act-tsca/epa-actions-reduce-risk-new
83. EU 2015/282. Commission Regulation (EU) 2015/282 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards the Extended one-generation reproductive toxicity study, February 2015
84. EU 528/2012. EU Biocidal Products Regulation (528/2012): EU, Regulation (EU) 528/2012 of the European Parliament and of the Council concerning the making available on the market and use of biocidal products, 22 May 2012.
85. EU B.26 Sub-chronic oral toxicity test repeated dose 90-day oral toxicity study in rodents. EC, Council Regulation (EC) 440/2008, laying down test methods pursuant to Regulation (EC) 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 30 May 2008.
86. EU B.35 Two-generation reproduction toxicity study. EC, Council Regulation (EC) 440/2008, laying down test methods pursuant to Regulation (EC) 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 30 May 2008.
87. EU B.53 Developmental neurotoxicity study. EU, Commission Regulation (EU) 900/2014 amending, for the purpose of its adaptation to technical progress, Regulation (EC) 440/2008 laying down test methods pursuant to Regulation (EC) 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 15 July 2014.
88. EU B.56 Extended one-generation reproductive toxicity study. EU, Commission Regulation (EU) 900/2014 amending, for the purpose of its adaptation to technical progress, Regulation (EC) 440/2008 laying down test methods pursuant to Regulation (EC) 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 15 July 2014.
89. EU, 2007. Corrigendum to Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Off. J. Eur. Union. L 136, 3–280, 29.5.2007.
90. European Union (2010) Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. OJEU L 276/33.
91. Everds, N.E., Snyder, P.W., Bailey, K.L., Bolon, B., Creasy, D.M., Foley, G.L., Rosol, T.J., Sellers, T., 2013. Interpreting stress responses during routine toxicity studies: a review of the biology, impact, and assessment. Toxicol. Pathol. 41, 560–614.
92. FDA (2013). United States Food and Drug Administration (FDA). Guidance document. Nonclinical safety evaluation of pediatric drug products. Silver Spring, MD: FDA.
93. FDA, 2004. Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.pdf.
94. Fegert I, Billington R, Botham P, Carney E, FitzGerald RE, Hanley T, Lewis R, Marty MS, Schneider S, Sheets LP, Stahl B, van Ravenzwaay B (2012) Feasibility of the extended one-generation reproductive toxicity study (OECD 443). Reprod Toxicol 34:331–339. doi.org/10.1016/j.reprotox.2012.05.096
95. Ferguson, S.A., Scallet, A.C., Flynn, K.M., Meredith, J.M., Schwetz, B.A., 2000. Developmental neurotoxicity of endocrine disrupters: focus on estrogens. Neurotoxicology 21, 947–56.
96. Fisher, R.A. Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh (1950).

97. Fleming A & National Centre for the Replacement, Refinement, and Reduction of Animals in Research (Great Britain). 2007. Zebrafish as an alternative model organism for disease modeling and drug discovery: implications for the 3Rs. NC3Rs [London].
98. Foster PMD (2014) Regulatory forum opinion piece: new testing paradigms for reproductive and developmental toxicity - the NTP Modified One Generation study and OECD 443. Toxicol Pathol XX:1-3. DOI: 10.1177/0192623314534920
99. Fraysse B, Mons R, Garric J. Development of a zebrafish 4-day embryo-larval bioassay to assess toxicity of chemicals. Ecotoxicology and Environmental Safety. 2006. 63:253–67.
100. Friedman JM. The principles of teratology: are they still true? Birth Defects Res A Clin Mol Teratol. 2010 Oct;88(10):766-8. doi: 10.1002/bdra.20697.
101. Fu W, Franco A, Trapp S. 2009. Methods for estimating the bioconcentration factor of ionizable organic chemicals. Environ Toxicol Chem 28: 1372-1379.
102. Gallavan, R.H. Jr, Holson, J.F., Stump, D.G., Knapp, J.F., Reynolds, V.L., 1999. Interpreting the toxicologic significance of alterations in anogenital distance: potential for confounding effects of progeny body weights. Reprod. Toxicol. 13, 383–390.
103. Garman RH, Abby AL, Kaufmann W, Auer RN, Bolon B (2015) Recommended Methods for Brain Processing and Quantitative Analysis in Rodent Developmental Neurotoxicity Studies. Toxicol Pathol XX: 1-29 DOI:10.1177/019262331559685
104. Gellert G, Heinrichsdorff J. Effect of age on the susceptibility of zebrafish eggs to industrial wastewater. Wat Res. 2001. 35(15):3754-7.
105. Genschow, E., Spielmann, H., Scholz, G., Seiler, A., Brown, N., Piersma, A., et al. (2002) The ECVAM international validation study on in vitro embryotoxicity tests: Results of the definitive phase and evaluation of prediction models. Altern Lab Anim 30:151–176.
106. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, Toppari J, Zoeller RT. Executive Summary to EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocr Rev. 2015 Dec;36(6):593-602. doi: 10.1210/er.2015-1093. Epub 2015 Sep 28.
107. Gore ER, Gower J, Kurali E, Sui JL, Bynum J, Ennulat D, Herzyk DJ. (2004) Primary antibody response to keyhole limpet hemocyanin in rat as a model for immunotoxicity evaluation. Toxicology 197(1):23-35
108. Green J, Wheeler JR. The use of carrier solvents in regulatory aquatic toxicology testing: Practical, statistical and regulatory considerations. Aquatic Toxicology. 2013. Volume 144–145:242–9.
109. Griggio, M.A., Luz, J., Gorgulho, A.A., Sucasas, C.M., 1997. The influence of food restriction during different periods of pregnancy. Int. J. Food Sci. Nutr. 48, 129–134.
110. Guidance of the Scientific Committee on a request from EFSA on the use of the benchmark dose approach in risk assessment. The EFSA Journal. 2009. 1150:1-72.
111. Guidelines for survival bleeding of mice and rats. http://oacu.od.nih.gov/ARAC/documents/
112. Gustafson 2008. Validation of zebrafish as a model for screening teratogenicity. 47th annual meeting of the society of toxicology. Toxicological sciences 102(S-1): 407.
113. Gustafson AL, Stedman DB, Ball J, Hillegass JM, Flood A, Zhang CX, Panzica-Kelly J, Cao J, Coburn A, Enright BP, Tornesi MB, Hetheridge M, Augustine-Rauch KA. Inter-laboratory assessment of a harmonized zebrafish developmental toxicology assay - Progress report on phase I. Reproductive toxicology. 2012. 33:155-64.
114. Hallare A, Nagel K, Kohler HR. Comparative embryotoxicity and proteotoxicity of three carrier solvents to zebrafish (Danio rerio) embryos. Ecotoxicology and environmental safety. 2006. 63:378-88.
115. Hallare AV, Kohler HR, Trienskorn R. Developmental toxicity and stress protein responses in zebrafish embryos after exposure to diclofenac and its solvent DMSO. Chemosphere. 2004. 56:659-66.
116. Hanley TR Jr, Watanabe PG (1985) Measurement of solid feed consumption patterns in neonatal rats by 141Ce-radiolabeled microspheres. Toxicol Appl Pharmacol 77(3):496-500
117. Harstad E, Andaya R, Couch J, Ding X, Liang X, Liederer BM, Messick K, Nguyen T, Schweiger M, Tarrant J, Zhong S, Dean B. Balancing Blood Sample Volume with 3Rs: Implementation and Best Practices for Small Molecule Toxicokinetic Assessments in Rats. ILAR J. 2016 Dec;57(2):157-165. doi: 10.1093/ilar/ilw023.
118. Henn K, Braunbeck T. Dechorionation as a tool to improve the fish embryo toxicity test (FET) with the zebrafish (Danio rerio). Comparative biochemistry and physiology, Part C. 2011. 153:91-8.
119. Hermsen SAB, van den Brandhof EJ, van der Ven LTM, Piersma AH. Relative embryotoxicity of two classess of chemicals in a modified zebrafish embryotoxicity test and comparison with their in vivo potencies. Toxicology in vitro. 2011. 25:745-53.

120. Herzyk DJ, Bussiere JL (2008) Immunotoxicology strategies for pharmaceutical safety assessment. John Wiley & Sons
121. Hill AJ, Bello SM, Prasch AL, Peterson RE, Heideman W. 2003. Water permeability and TCDD-induced edema in zebrafish early-life stages. Toxicological sciences 78(1): 78-87.
122. Holmes AM, Creton S, Chapman K. Working in partnership to advance the 3Rs in toxicity testing. Toxicology. 2010 Jan 12;267(1-3):14-9. doi: 10.1016/j.tox.2009.11.006. Epub 2009 Nov 10.
123. Holsapple, M.P., Burns-Naas, L.A., Hastings, K.L., Ladics, G.S., Lavin, A.L., Makris, S.L., Yang, Y., Luster, M.I., 2005. A proposed testing framework for developmental immunotoxicology (DIT). Toxicol Sci. 83, 18–24.
124. Holsapple, M.P., O'Lone, R., 2012. Juvenile immunotoxicology. Toxicol. Pathol. 40:248–254.
125. House, R.V., 1999. Theory and practice of cytokine assessment in immunotoxicology. Methods 19, 17–27.
126. Huang H, Huang C, Wang L, Ye X, Bai C, Simonich MT, Tanguay RL, Dong Q. Toxicity, uptake kinetics and behavior assessment in zebrafish embryos following exposure to perfluorooctanesulphonicacid (PFOS). Aquat Toxicol. 2010. 98:139-47.
127. Hurtt ME, Cappon GD, Browning A. Proposal for a tiered approach to developmental toxicity testing for veterinary pharmaceutical products for food-producing animals. Food Chem Toxicol; 41; 611-9 (2003).
128. ICH M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals, EMA/CPMP/ICH/286/1995, December 2009.
129. ICH S3A Note for Guidance on Toxicokinetics: The Assessment of Systemic Exposure in Toxicity Studies, October 1994.
130. ICH S5 (R2) harmonized tripartite guideline. Detection of toxicity to reproduction for medicinal products & toxicity to male fertility S5. Parent guideline dated 24 June 1993. Addendum dated 9 November 2000 incorporated in November 2005.
131. ICH S5 (R2): International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline on Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility, parent guideline dated 24 June 1993 (addendum dated 9 November 2000 incorporated in November 2005), Section 4.1.3. "Study for effects on embryo-fetal development".
132. ICH S5(R3) Draft Concept Paper: Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility (Revision of the ICH S5(R2) Guideline: Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility). Issued 27 January 2014.
133. ILSI/HESI (International Life Sciences Institute/Health and Environmental Science Institute), 2001. Developing Strategies for Agricultural Chemical Safety Evaluation: A Report from April 22–23, 2001 workshop. September 2001, Washington, DC.
134. IPCS (2002). Global assessment of the state-of-the-science of endocrine disruptors. International Programme on Chemical Safety, World Health Organization and United Nations Environment Programme, Geneva
135. Ito T, Handa H. Deciphering the mystery of thalidomide teratogenicity. Congenit Anom (Kyoto). 2012 Mar;52(1):1-7. doi: 10.1111/j.1741-4520.2011.00351.x.
136. Janer G, Hakkert BC, Slob W, Vermeire T, Piersma AH. A retrospective analysis of the two-generation study: what is the added value of the second generation? Reprod Toxicol 2007;24:97–102.
137. Janer G, Slob W, Hakkert BC, Vermeire T, Piersma AH. A retrospective analysis of developmental toxicity studies in rat and rabbit: what is the added value of the rabbit as an additional test species? Regul Toxicol Pharmacol. 50:206–217 (2008).
138. Jomaa B, Hermsen SAB, Kessels MY, van den Berg JHJ, Peijnenburg AACM, Aarts JMMJG, Piersma AH, Rietjens IMCM. Developmental toxicity of thyroid-active compounds in a zebrafish embryotoxicity test. ALTEX. 2014;31(3):303-17.
139. Jonsson, O., Palma Villar, R., Nilsson, L.B., Eriksson, M., and Königsson, K. (2012). Validation of a bioanalytical method using capillary microsampling of 8 µl plasma samples: application to a toxicokinetic study in mice. Bioanalysis 4, 1989 – 1998. doi:10.4155/bio.12.177
140. Jonsson, O., Palma Villar, R., Nilsson, L.B., Norsten-Höög, C., Brogren, J., Eriksson, M., Königsson, K., and Samuelsson, A. (2012). Capillary microsampling of 25µl blood for the determination of toxicokinetic parameters in regulatory studies in animals. Bioanalysis 4, 661-674. doi: 10.4155/bio.12.25.
141. Judson R, Houck K, Martin M, Knudsen T, Thomas RS, Sipes N, Shah I, Wambaugh J, Crofton K. In vitro and modelling approaches to risk assessment from the U.S. Environmental Protection Agency ToxCast programme. Basic Clin Pharmacol Toxicol. Jul;115(1):69-76. doi: 10.1111/bcpt.12239 (2014).

142. Kais B, Schneider KE, Keiter S, Henn K, Ackermann C, Braunbeck T. DMSO modifies the permeability of the zebrafish (Danio rerio) chorion-Implications for the fish embryo test (FET). Aquatic Toxicology. 2013. Volume 140-141:229-38.
143. Kalland, T., Strand, Ø., Forsberg, J.G., 1979. Long-term effects of neonatal estrogen treatment on mitogen responsiveness of mouse spleen lymphocytes. J. Natl. Cancer Inst. 63, 413–421.
144. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Developmental dynamics. 1995. 203:253-310.
145. Klüver H, Barrera E (1953) A method for the combined staining of cells and fibers in the nervous system. J Neuropathol Exp Neurol 12:400.
146. Knapen D, Vergauwen L, Villeneuve DL, Ankley GT. The potential of AOP networks for reproductive and developmental toxicity assay development. Reprod Toxicol. Aug 15;56:52-5. doi: 10.1016/j.reprotox.2015.04.003 (2015).
147. Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh AV. Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. Reprod Toxicol. 28:209–219 (2009).
148. Könemann H, van Leeuwen K 1980. Toxicokinetics in fish: accumulation and elimination of six chlorobenzenes by guppies. Chemosphere 9: 3-19.
149. Korfmacher, W., Luo, Y, Ho., S., Sun, W., Shen, L., Wang, J., Wu, Z., Guo, Y., Snow, G., and O'Shea, Th. (2015). Utility of capillary microsampling for rat pharmacokinetic studies: Comparison of tail-vein bleed to jugular vein cannula sampling. J. Pharmacol. Toxicol. Meth. 76, 7-14. doi: 10.1016/j.vascn.2015.07.001
150. Kortenkamp A, Martin O, Faust M, Evans R, McKinlay R, Orton F, Rosivatz E. (2011) State of the Art Assessment of Endocrine Disruptors. Final Report, Project Contract Number 070307/2009/550687/SER/D3. Available at: http://ec.europa.eu/environment/chemicals/endocrine/pdf/sota_edc_final_report.pdf
151. Krewski D, Acosta D Jr, Andersen M, Anderson H, Bailar JC 3rd, Boekelheide K, Brent R, Charnley G, Cheung VG, Green S Jr, Kelsey KT, Kerkvliet NI, Li AA, McCray L, Meyer O, Patterson RD, Pennie W, Scala RA, Solomon GM, Stephens M, Yager J, Zeise L. Toxicity testing in the 21st century: a vision and a strategy. J Toxicol Environ Health B Crit Rev. Feb;13(2-4):51-138. doi: 10.1080/10937404.2010.483176 (2010).
152. Kroese ED, Bosgra S, Buist HE, Lewin G, van der Linden SC, Man HY, Piersma AH, Rorije E, Schulpen SH, Schwarz M, Uibel F, van Vugt-Lussenburg BM, Wolterbeek AP, van der Burg B. Evaluation of an alternative in vitro test battery for detecting reproductive toxicants in a grouping context. Reprod Toxicol. Aug 1;55:11-9. doi: 10.1016/j.reprotox.2014.10.003 (2015).
153. Kruskal, W.H., Wallis, W.A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47 583-621 (1952).
154. Kuper, C.F., Harleman, J.H., Richter-Reichelm, H.B., Vos, J.G., 2000. Histopathologic approaches to detect changes indicative of immunotoxicity. Toxicol. Pathol. 28, 454–466.
155. Lai DY. Rodent carcinogenicity of peroxisome proliferators and issues on human relevance. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 2004 May;22(1):37-55.
156. Lammer E, Kamp HG, Hisgen V. Development of a flow-through system for the fish embryo toxicity test (FET) with the zebrafish (Danio rerio). Toxicology in vitro. 2009. 23:1436-42.
157. Landreth KS (2002) Critical windows in development of the rodent immune system. Hum Exp Toxicol 21:493-498.
158. Lee HY, Inselman AL, Kanungo J, Hansen DK. Alternative models in developmental toxicology. Syst Biol Reprod Med. 2012 Feb;58(1):10-22. doi: 10.3109/19396368.2011.648302.
159. Lee KJ, Nallathamby PD, Browning LM, Osgood CJ, Xu XN. In vivo imaging of transport and biocompatibility of single silver nanoparticles in early development of zebrafish embryos. ACS Nano. 2007. 28, 1(2):133–43.
160. Lenz W. (1961). Klinische Missbildungen nach Medikament-Einnahme wahrend der Graviditat? Dtsch Med Wochensch 86: 2555–2556.
161. Levine, T.E., Butcher, R.E., 1990. Workshop on the qualitative and quantitative comparability of human and animal developmental neurotoxicity, work group IV report: triggers for developmental neurotoxicity testing. Neurotoxicol. Teratol. 12, 281–284.
162. Luebke, R.W., Chen, D.H., Dietert, R., Yang, Y., King, M., Luster, M.I., 2006. The comparative immunotoxicity of five selected compounds following developmental or adult exposure. J. Toxicol. Environ. Health B 9, 1–26.

163. Luo, Y, Korfmacher, W., Ho., S., Shen, L., Wang, J., Wu, Z., Guo, Y., Snow, G., and O'Shea, Th. (2015). Evaluation of two blood sampling approaches for drug discovery PK studies in rats. Bioanalysis 7, 2345-2359. doi: 10.4155/bio.15.126
164. Luster, M.I., Portier, C., Pait, D.G., White, K.L. Jr, Gennings, C., Munson, A.E., Rosenthal, G.J., 1992. Risk assessment in immunotoxicology. I. Sensitivity and predictability of immune tests. Fundam. Appl. Toxicol. 18, 200–210.
165. Maes J, Verlooy L, Buenafe O. Evaluation of 14 organic solvents and carriers for screening applications in zebrafish embryos and larvae. Plos One. 2012. 7(10):1-9.
166. Mandrell D, Truong L, Jephson C, Sarker MR, Moore A, Lang C, Simonich MT, Tanguay RL. Automated zebrafish chorion removal and single embryo placement: optimizing throughput of zebrafish developmental toxicity screens. J Lab Autom. 2012. 17(1):66–74.
167. Mann, H. B., Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18 (1), 50–60 (1947).
168. Martin MT, Mendez E, Corum DG, Judson RS, Kavlock RJ, Rotroff DM, Dix DJ. Profiling the reproductive toxicity of chemicals from multi-generation studies in the toxicity reference database. Toxicol Sci 2009;110:181–190.
169. Marty MS, Neal BH, Zablotny CL, Yano BL, Andrus AK, Woolhiser MR, Boverhof DR, Saghir SA, Perala AW, Passage JK, Lawson MA, Bus JS, Lamb JS, Hammond L (2013) An F1-Extended One-Generation Reproductive Toxicity Study in Crl:CD(SD) rats with 2,4-Dichlorophenoxyacetic acid. Toxicol Sci 136(2):527–547. doi:10.1093/toxsci/kft213
170. Marzo M, Kulkarni S, Manganaro A, Roncaglioni A, Wu S, Barton-Maclaren TS, Lester C, Benfenati E. Integrating in silico models to enhance predictivity for developmental toxicity. Toxicology. Aug 31;370:127-137. doi: 10.1016/j.tox.2016.09.015 (2016).
171. McBride WG. (1961). Thalidomide and congenital abnormalities. Lancet 2: 1358.
172. Meek ME, Bucher JR, Cohen SM, Dellarco V, Hill RN, Lehman-McKeeman LD, Longfellow DG, Pastoor T, Seed J, Patton DE. A framework for human relevance analysis of information on carcinogenic modes of action. Crit Rev Toxicol. 2003;33(6):591-653.
173. Melching-Kollmuß, S., Fussell, K.C., Buesen, R., Dammann, M., Schneider, S., Tennekes, H., van Ravenzwaay, B., 2014. Anti-androgenicity can only be evaluated using a weight of evidence approach. Regul. Toxicol. Pharmacol. 68, 175–192.
174. Merrick BA, Paules RS, Tice RR. Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. Int J Biotechnol. 14(1):7-27 (2015).
175. Miller, R.G. Simultaneous Statistical Inference, Springer Verlag, New York (1981).
176. Mitchard T, Kirk S, Grant C, Stewart J. Investigation into the effect of microsampling on mouse fetuses and pregnant mice in the embryofetal development study design. Reprod Toxicol. 2017 Jan;67:140-145. doi: 10.1016/j.reprotox.2016.12.006. Epub 2016 Dec 30.
177. Moore, N., Bremer, S., Carmichael, N., Daston, G., Dent, M., Gaoua-Chapelle, W., Hallmark, M., Hartung, T., Holzum, B., Hübel, U., Meisters, M.L., Schneider, S., van Ravenzwaay, B., Hennes, C., 2009. A modular approach to the extended one-generation reproduction toxicity study. Altern. Lab. Anim., 37, 219–225.
178. Moore, N.P., Boogaard, P.J., Bremer, S., Buesen, R., Edwards, J., Fraysse, B., Hallmark, N., Hemming, H., Langrand-Lerche, C., McKee, R.H., Meisters, M.L., Parsons, P., Politano, V., Reader, S., Ridgway, P., Hennes, C., 2013. Guidance on classification for reproductive toxicity under the Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Crit Rev Toxicol. 43, 850–891.
179. Myers, D.P., Willoughby, C.R., Bottomley, A.M., Buschmann, J., 2008. An analysis of the results from Two-Generation Reproduction Toxicity Studies to assess the value of mating animals of the second (F1) generation for the detection of adverse treatment-related effects on reproductive performance. Reprod. Toxicol. 26, 47–50.
180. NC3Rs decision tree for microsampling https://www.nc3rs.org.uk/sites/default/files/Microsampling/microsampling%20decision%20tree.pdf (accessed April 28, 2017).
181. Nillson LB, Ahnoff M, Jonsson O. (2013). Capillary microsampling in the regulatory environment: validation and use of bioanalytical capillary microsampling methods. Bionanalysis 5, 731-738. doi: 10.4155/bio.13.27
182. Niu X, Beekhuijzen M, Schoonen W, Emmen H, Wenker M. Effects of capillary microsampling on toxicological endpoints in juvenile rats. Toxicological Sciences, 1-9 (2016).

183. NRC, Committee on Developmental Toxicology, Board on Environmental Studies and Toxicology, National Research Council. Scientific frontiers in developmental toxicology and risk assessment. Washington, DC: National Academy Press; 2000.
184. NRC, US National Academy of Sciences report on Toxicity Testing in the 21st Century – a Vision and a Strategy. National Research Council. The National Academies Press, Washington, D.C. 2007.
185. OECD (1998). Report of the First Meeting of the OECD Endocrine Disruptor Testing and Assessment (EDTA) Task Force, 10th-11th March 1998, ENV/MC/CHEM/RA(98)5.
186. OECD (2006). Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-day Oral Toxicity Study in Laboratory Rats. Series on Testing and Assessment No 59, ENV/JM/MONO(2006)26.
187. OECD (2015). Feasibility Study for Minor Enhancements of TG 421/422 with ED Relevant Endpoints. Environment, Health and Safety Publications, Series on Testing and Assessment (No. 217), Organisation for Economic Cooperation and Development, Paris.
188. OECD 236, OECD guideline for testing of chemicals No. 236: Fish embryo acute toxicity (FET) test. Adopted 26 July 2013.
189. OECD 407. OECD guideline for testing of chemicals No. 407: Repeated dose 28-day oral toxicity study in rodents. October 2008.
190. OECD 408. OECD guideline for testing of chemicals No. 408: Repeated dose 90-day oral toxicity study in rodents. September 1998
191. OECD 413, 2009. OECD Guideline for the Testing of Chemicals. Test Guideline 413: Subchronic Inhalation Toxicity: 90-Day Study. Organisation for Economic Co-operation and Development, Paris.
192. OECD 414. Guideline for the testing of chemicals. No. 414. Prenatal developmental toxicity study. January 2001. (Previous version in 1981).
193. OECD 415. Guideline for the testing of chemicals. No. 415. One-generation reproduction toxicity study. May 1983.
194. OECD 416. Guideline for the testing of chemicals. No. 416. Two generation reproduction toxicity study. January 2001. (Previous version in 1983).
195. OECD 420, 2001. OECD Guideline for the Testing of Chemicals. Test Guideline 420: Acute Oral Toxicity – Fixed Dose Procedure. Organisation for Economic Co-operation and Development, Paris.
196. OECD 421. OECD guideline for testing of chemicals No. 421: Reproduction/developmental toxicity screening test. July 2016. (Previous versions in 1995 and 2015).
197. OECD 422. OECD guideline for testing of chemicals No. 422: Combined repeated dose toxicity study with the reproduction/developmental toxicity screening test. July 2016. (Previous versions in 1996 and 2015).
198. OECD 423, 2001. OECD Guideline for the Testing of Chemicals. Test Guideline 423: Acute Oral Toxicity – Acute Toxic Class Method. Organisation for Economic Co-operation and Development, Paris.
199. OECD 425, 2008. OECD Guideline for the Testing of Chemicals. Test Guideline 425: Acute Oral Toxicity – Up-and-Down-Procedure (UDP). Organisation for Economic Co-operation and Development, Paris.
200. OECD 426. OECD guideline for testing of chemicals No. 426: Developmental neurotoxicity study. October 2007
201. OECD 440, 2007. OECD Guideline for the Testing of Chemicals. Test Guideline 440: Uterotrophic Bioassay in Rodents: A short-term Screening Test for Oestrogenic Properties. Organisation for Economic Co-operation and Development, Paris.
202. OECD 441, 2009. OECD Guideline for the Testing of Chemicals. Test Guideline 441: Hershberger Bioassay in Rats: A Short-term Screening Assay for (Anti)Androgenic Properties. Organisation for Economic Co-operation and Development, Paris.
203. OECD 443. OECD guideline for the testing of chemicals No. 443. Extended one generation reproductive toxicity study. October 2012 (Previous version in 2011).
204. OECD GD 117. Guidance Document no. 117 on the current implementation of internal triggers in test guideline 443 for an extended one generation reproductive toxicity study, in the United States and Canada. ENV/JM/MONO(2011)21; July 2011.
205. OECD GD 150 (2012) Environment, Health and Safety Publications, Series on Testing and Assessment No. 150, Guidance document on standardised test guidelines for evaluating chemicals for endocrine disruption, Environment Directorate Organisation for Economic Co-operation and Development
206. OECD GD 151. Guidance Document no. 151 supporting OECD test guideline 443 on the extended one generation reproductive toxicity test. ENV/JM/MONO(2013)10. July 2013.

207. OECD GD 32, 2002. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. OECD Series on Testing and Assessment: No. 32. Organisation for Economic Co-operation and Development, Paris.
208. OECD GD 35, 2002. Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies. OECD Series on Testing and Assessment: No. 35. Organisation for Economic Co-operation and Development, Paris.
209. OECD GD 43, 2008. Guidance Document on Mammalian Reproductive Toxicity Testing and Assessment. Series on Testing and Assessment: No. 43. Organisation for Economic Co-operation and Development, Paris.
210. OECD Guidelines for Testing of Chemicals No 210: Fish, Early-life Stage Toxicity Test. Adopted 26 July 2013
211. OECD Guidelines for Testing of Chemicals No. 229: Fish Short Term Reproduction Assay. Adopted 02 October 2012
212. OECD Guidelines for Testing of Chemicals No. 230: 21-day Fish Assay. Adopted 08 September 2009
213. OECD Guidelines for Testing of Chemicals No.231: Amphibian Metamorphosis Assay. Adopted 08 September 2009
214. OECD Guidelines for Testing of Chemicals No.234: Fish Sexual Development Test. Adopted 28 July 2011
215. OECD No. 176. Report on the retrospective analysis of 2-generation reprotoxicity data. ENV/JM/MONO(2012)21. July 2012.
216. OECD Series on Testing and Assessment No.23. Guidance document on aquatic toxicity testing of difficult substances and mixtures. Adopted 14 December 2000.
217. OECD, 2012. The adverse outcome pathway for skin sensitisation initiated by covalent binding for proteins. Part 1. Scientific evidence OECD environment, health and safety publications series on testing and assessment.
218. OECD, 2014. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Part 1: Scientific Evidence. Series on Testing and Assessment: No. 168. Organisation for Economic Co-operation and Development, Paris.
219. Owens W, Ashby J, Odum J, Onyon L (2003) The OECD program to validate the rat uterotrophic bioassay. Phase 2: dietary phytoestrogen analyses. Environ Health Perspect 111(12):1559-1567
220. Padilla S, Corum D, Padnos B, Hunter DL, Beam A, Houck KA, Sipes N, Kleinstreuer N, Knudsen T, Dix DJ, Reif DM. Zebrafish developmental screening of the ToxCast™ Phase I chemical library. Reprod Toxicol. 2012. 33(2):174-87.
221. Palmer AK and Ulbrich BC (1997) The Cult of Culling. Fundamental and Applied Toxicology 38:7-22
222. Panzica-Kelly JM, Zhang CX, Danberry TL, Flood A, DeLan JW, Brannen KC, Augustine-Rauch KA. Morphological Score Assignment Guidelines for the Dechorionated Zebrafish Teratogenicity Assay. Birth Defects Research (Part B). 2010. 89:382-95.
223. Paxinos G, Watson C (1998) The Rat Brain in Stereotaxic Coordinates. 4th Edition. Academic Press, New York
224. Pelka KE, Henn K, Keck A, Sapel B, Braunbeck T. Size does matter - Determination of the critical molecular size for the uptake of chemicals across the chorion of zebrafish (Danio rerio) embryos. Aquat Toxicol. 2017 Apr;185:1-10. doi: 10.1016/j.aquatox.2016.12.015. Epub 2016 Dec 21.
225. Peterson RT, Fishman MC. 2011. Designing zebrafish chemical screens. Methods Cell Biol 105: 525-541.
226. Piersma AH, Ezendam J, Luijten M, Muller JJ, Rorije E, van der Ven LT, van Benthem J. A critical appraisal of the process of regulatory implementation of novel in vivo and in vitro methods for chemical hazard and risk assessment. Crit Rev Toxicol. Nov;44(10):876-94. doi: 10.3109/10408444.2014.940445 (2014).
227. Piersma AH, Hernandez LG, van Benthem J, Muller JJ, van Leeuwen FX, Vermeire TG, van Raaij MT. Reproductive toxicants have a threshold of adversity. Crit Rev Toxicol. 2011 Jul;41(6):545-54. doi: 10.3109/10408444.2011.554794. Epub 2011 May 24.
228. Piersma AH, Janer G, Wolterink G. Quantitative extrapolation of In vitro whole embryo culture embryotoxicity data to developmental toxicity in vivo using the benchmark dose approach. Toxicological sciences. 2008. 101():91-100.
229. Piersma AH, Rorije E, Beekhuijzen MEW, Cooper R, Dix DJ, Heinrich-Hirsch B, Martin MT, Mendez E, Muller A, Paparella M, Ramsingh D, Reaves E, Ridgway P, Schenk E, Stachiw L, Ulbrich B, Hakkert BC (2011) Combined retrospective analysis of 498 rat multi-generation reproductive toxicity studies: on the

impact of parameters related to F1 mating and F2 offspring. Reprod Toxicol 31:392–401. doi: 10.1016/j.reprotox.2010.11.013
230. Piersma AH. Alternative methods for developmental toxicity testing. Basic Clin Pharmacol Toxicol 2006; 98(5):427-431.
231. Piersma, A.H., Tonk, E.C.M., Makris, S.L., Crofton, K.M., Dietert, R.R., van Loveren, H., 2012. Juvenile toxicity testing protocols for chemicals. Reprod.Toxicol. 34, 482–486.
232. Powles-Glover N, Kirk S, Jardine L, Clubb S, Stewart J. (2014). Assessment of haematological and clinical pathology effects of blood microsampling in suckling and weaned juvenile rats. Regul. Toxicol. Pharmacol. 69, 425-433. doi: 10.1016/j.yrtph.2014.05.006.
233. Powles-Glover N, Kirk S, Wilkinson C, Robinson S, Stewart J. (2014). Assessment of toxicological effects of blood microsampling in the vehicle dosed adult rat. Regul. Toxicol. Pharmacol. 68, 325-331. doi: 10.1016/j.yrtph.2014.01.001.
234. Pype C, Verbueken E, Saad MA, Bars C, Van Ginneken CJ, Knapen D, Van Cruchten SJ. Antioxidants reduce reactive oxygen species but not embryotoxicity in the metabolic Danio rerio test (mDarT). Reprod Toxicol. 2017 Jun 27. pii: S0890-6238(17)30174-0. doi: 10.1016/j.reprotox.2017.06.132. [Epub ahead of print].
235. Raffaele KC, Rowland J, May B, Makris SL, Schumacher K, Scarano LJ. (2010) The use of developmental neurotoxicity data in pesticide risk assessments. Neurotoxicol Teratol 32:563-572. doi:10.1016/j.ntt.2010.04.053
236. Rawson DM, Zhang T, Kalicharan D, Jongebloed WL. 2000. Field emission scanning electron microscopy and transmission electron microscopy studies of the chorion, plasma membrane and syncytial layers of the gastrula-stage embryo of the zebrafish Brachydanio rerio: a consideration of the structural and functional relationships with respect to cryoprotectant penetration. Aquaculture Research 31: 325-336.
237. Reuter U, Heinrich-Hirsch B, Hellwig J, Holzum B, Welsch F. Evaluation of OECD screening tests 421 (reproduction/developmental toxicity screening test) and 422 (combined repeated dose toxicity study with the reproduction/developmental toxicity screening test). Regul Toxicol Pharm 2003;38:17–26.
238. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol. Aug 15;29(8):1225-51. doi: 10.1021/acs.chemrestox.6b00135 (2016).
239. Rodent_Bleeding.pdf
240. Romero, A., Villamayor, F., Grau, M.T., Sacristán, A., Ortiz, J.A., 1992. Relationship between fetal weight and litter size in rats: application to reproductive toxicology studies. Reprod. Toxicol. 6, 453–456.
241. Rorije E, Muller A, Beekhuijzen ME, Hass U, Heinrich-Hirsch B, Paparella M, Schenk E, Ulbrich B, Hakkert BC, Piersma AH (2011) On the impact of second generation mating and offspring in multi-generation reproductive toxicity studies on classification and labelling of substances in Europe. Regul Toxicol Pharmacol 61: 251-60. doi: 10.1016/j.yrtph.2011.08.005.
242. Russell, W.M.S. and Burch, R.L., (1959). The Principles of Humane Experimental Technique, Methuen, London. ISBN 0900767782.
243. S3A Implementation Working Group. ICH S3A Guideline: Note for Guidance on Toxicokinetics: The Assessment of Systemic Exposure in Toxicity Studies Questions and Answers, 19 January 2016 [Internet]. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S3A/ICH_S3A_draft_Q_A_Step1-29Jan2016.pdf
244. Sadler AM, Bailey SJ. (2013). Validation of a refined technique for taking repeated blood samples from juvenile and adult mice. Lab. Anim. 47, 316-319. doi: 10.1177/0023677213494366.
245. Saghir SA. Rethinking guideline toxicity testing. Regul Toxicol Pharmacol. Jul; 72(2):423-8 (2015).
246. Salewski, E. Färbemethode zum makroskopischen Nachweis von Implantationsstellen am Uterus der Ratte. [Staining method for a macroscopic test for implantation sites in the uterus of the rat]. Naunyn - Schmiedebergs Archiv für Experimentelle Pathologie und Pharmakologie, 247, 367 (1964).
247. Schreiber R, Altenburger R, Paschke A, Schüürmann G, Küster E. 2009. A novel in vitro system for the determination of bioconcentration factors and the internal dose in zebrafish (Danio rerio) eggs. Chemosphere 77: 928-933.
248. Selderslaghs IWT, Blust R, Witters HE. Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. Reproductive toxicology. 2012. 33:142-54.

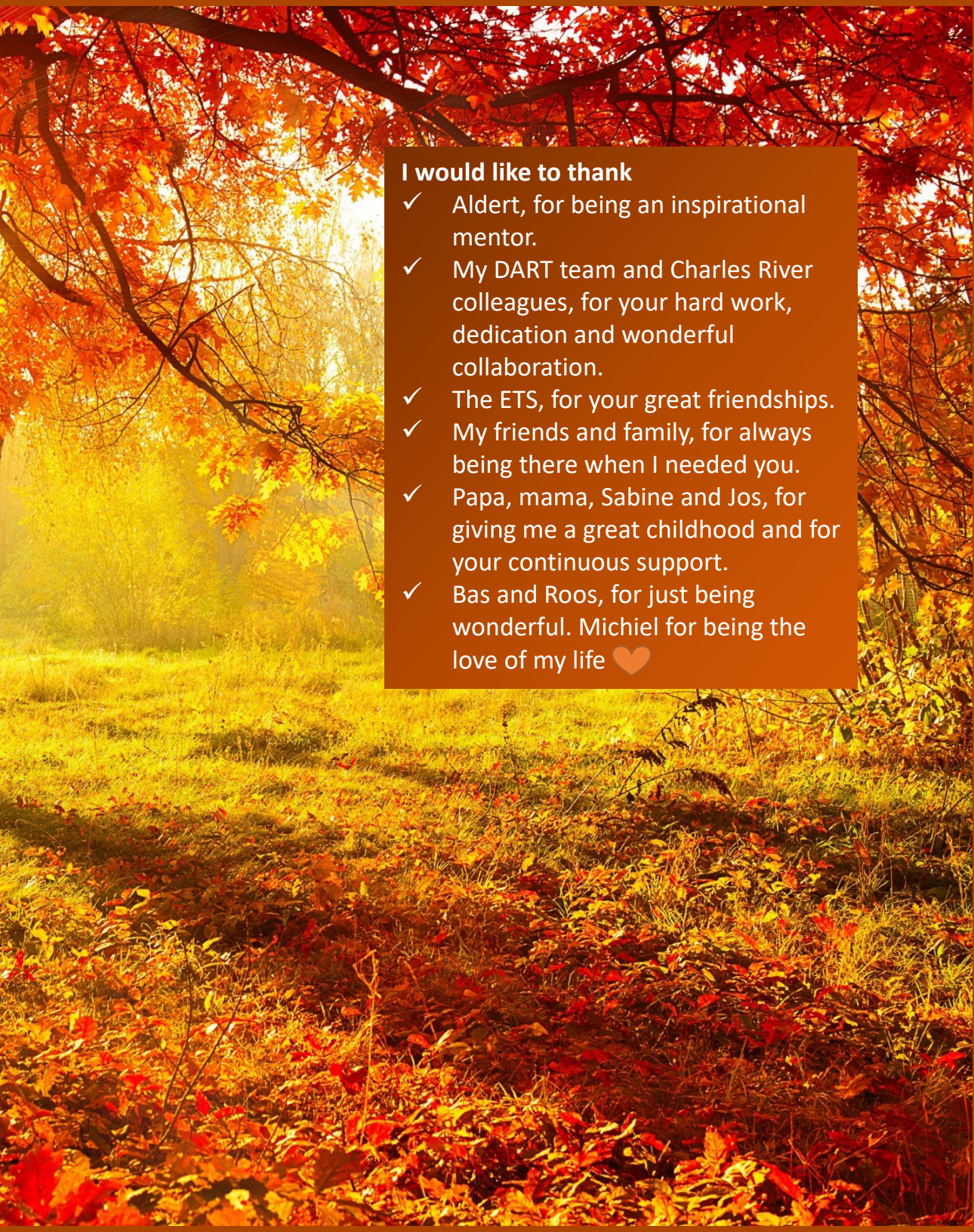
249. Selderslaghs IWT, Rompay AR, de Coen W. Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo. Reproductive toxicology. 2009. 28:308-20.
250. Selevan SG, Kimmel CA, Mendola P. Identifying critical windows of exposure for children's health. Environ Health Perspect. 2000 Jun;108 Suppl 3:451-5.
251. Shirley, E., 1977. The analysis of organ weight data. Toxicology 8, 13–22.
252. Slob W. Dose-response modeling of continuous endpoints. Toxicological science. 2002. 66:298-312.
253. Smialowicz, R.J., Riddle, M.M., Rogers, R.R., Rowe, D.G., Luebke, R.W., Fogelson, L.D., Copeland, C.B., 1988. Immunologic effects of perinatal exposure of rats to dioctyltin dichloride. J. Toxicol. Environ. Health 25, 403–422.
254. Smith K, Oostingh G, Mayer P. Passive dosing for producing defined and constant exposure of hydrophobic organic compounds during in vitro toxicity tests. Chem. Res. Toxicol. 2010. 23(1):55–65.
255. Sogorb MA, Pamies D, de Lapuente J, Estevan C, Estévez J, Vilanova E. An integrated approach for detecting embryotoxicity and developmental toxicity of environmental contaminants using in vitro alternative methods. Toxicol Lett. Oct 15;230(2):356-67. doi: 10.1016/j.toxlet.2014.01.037 (2014).
256. Soucy, G., Boivin, G., Labrie, F., Rivest, S., 2005. Estradiol is required for a proper immune response to bacterial and viral pathogens in the female brain. J. Immunol. 174, 6391–6398.
257. Sparrow SS, Robinson S, Bolam S, Bruce C, Danks A, Everett D, Fulcher S, Hill RE, Palmer H, Scott EW, Chapman KL. 2011. Opportunities to minimise animal use in pharmaceutical regulatory general toxicology: a cross-company review. Regul Toxicol Pharmacol 61(2):222–229.
258. Steinberg P. 2013. High-Throughput Screening Methods in Toxicity Testing. Wiley: ISBN: 978-1-118-06563-1.
259. Strähle U, Scholz S, Geisler R, Greiner P, Hollert H, Rastegar S, Schumacher A, Selderslaghs I, Weiss C, Witters H, Braunbeck T. Zebrafish embryos as an alternative to animal experiments - a commentary on the definition of the onset of protected life stages in animal welfare regulations. Reproductive Toxicology. 2012. 33(2):128-32.
260. Stuckhardt, J.L.; Poppe, S.M. Fresh visceral examination of rat and rabbit fetuses used in teratogenicity testing. Teratogenesis, Carcinogenesis and Mutagenesis,4, 181-188 (1984).
261. Sullivan FM. Significance of Excursions of Intake above the Acceptable Daily Intake: Effect of Time and Dose in Developmental Toxicology. Regulatory Toxicology and Pharmacology 30, S94–S98 (1999)
262. Taub DD. Neuroendocrine interactions in the immune system. Cell Immunol. 2008 Mar-Apr;252(1-2):1-6. doi: 10.1016/j.cellimm.2008.05.006. Epub 2008 Jul 10.
263. Teixidó E, Piqué E, Gómez-Catalán J, Llobet JM. Assessment of developmental delay in the zebrafish embryo teratogenicity assay. Toxicol In Vitro. 2013. 27(1):469-78.
264. Testai E, Galli CL, Dekant W, Marinovich M, Piersma AH, Sharpe RM. A plea for risk assessment of endocrine disrupting chemicals. Toxicology 2013;314: 51–9.
265. Theunissen PT, Beken S, Beyer BK, Breslin WJ, Cappon GD, Chen CL, Chmielewski G, De Schaepdrijver L, Enright B, Foreman JE, Harrouk W, Hew KW, Hoberman AM, Hui JY, Knudsen TB, Laffan SB, Makris SL, Martin M, McNerney ME, Siezen CL, Stanislaus DJ, Stewart J, Thompson KE, Tornesi B, Van der Laan JW, Weinbauer GF, Wood S, Piersma AH. Comparison of rat and rabbit embryo-fetal developmental toxicity data for 379 pharmaceuticals: on the nature and severity of developmental effects. Critical Reviews in Toxicology. Vol. 46 No. 10, 900-910 (2016).
266. Theunissen PT, Beken S, Beyer BK, Breslin WJ, Cappon GD, Chen CL, Chmielewski G, De Schaepdrijver L, Enright B, Foreman JE, Harrouk W, Hew KW, Hoberman AM, Hui JY, Knudsen TB, Laffan SB, Makris SL, Martin M, McNerney ME, Siezen CL, Stanislaus DJ, Stewart J, Thompson KE, Tornesi B, Van der Laan JW, Weinbauer GF, Wood S, Piersma AH. Comparing rat and rabbit embryo-fetal developmental toxicity data for 379 pharmaceuticals: on systemic dose and developmental effects. Critical Reviews in Toxicology, DOI: 10.1080/10408444.2016.1224808 (2016).
267. Tice RR, Austin CP, Kavlock RJ, Bucher JR. Improving the human hazard characterization of chemicals: a Tox21 update. Environ Health Perspect. Jul;121(7):756-65. doi: 10.1289/ehp.1205784 (2013).
268. Tollefsen KE, Scholz S, Cronin MT, Edwards SW, de Knecht J, Crofton K, Garcia-Reyero N, Hartung T, Worth A, Patlewicz G. Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul Toxicol Pharmacol. Dec;70(3):629-40. doi: 10.1016/j.yrtph.2014.09.009 (2014).

269. Tonk, E.C.M., de Groot, D.M., Penninks, A.H., Waalkens-Berendsen, I.D., Wolterbeek, A.P., Piersma, A.H., van Loveren, H., 2011. Developmental immunotoxicity of di-n-octyltin dichloride (DOTC) in an extended one-generation reproductive toxicity study. Toxicol. Lett. 204, 156–63.
270. Tonk, E.C.M., de Groot, D.M., Penninks, A.H., Waalkens-Berendsen, I.D., Wolterbeek, A.P., Slob, W., Piersma, A.H., van Loveren, H., 2010. Developmental immunotoxicity of methylmercury: the relative sensitivity of developmental and immune parameters. Toxicol. Sci. 117, 325–35.
271. Tonk, E.C.M., de Groot, D.M., Wolterbeek, A.P., Penninks, A.H., Waalkens-Berendsen, I.D., Piersma, A.H., van Loveren, H., 2013. Developmental immunotoxicity of ethanol in an extended one-generation reproductive toxicity study. Arch. Toxicol. 87, 323–35.
272. Tonk, E.C.M., Verhoef, A., de la Fonteyne, L.J., Waalkens-Berendsen, I.D., Wolterbeek, A.P., van Loveren, H., Piersma, A.H., 2011. Developmental immunotoxicity in male rats after juvenile exposure to di-n-octyltin dichloride (DOTC). Reprod. Toxicol. 32, 341–8.
273. Tonk, E.C.M., Verhoef, A., Gremmer, E.R., van Loveren, H., Piersma, A.H., 2012. Relative sensitivity of developmental and immune parameters in juvenile versus adult male rats after exposure to di(2-ethylhexyl) phthalate. Toxicol. Appl. Pharmacol. 260, 48–57.
274. Tonk, E.C.M., Verhoef, A., Gremmer, E.R., van Loveren, H., Piersma, A.H., 2013. Developmental immunotoxicity in male rats after juvenile exposure to ethanol. Toxicology 309, 91–99.
275. Tonk, E.C.M., Verhoef, A., Gremmer, E.R., van Loveren, H., Piersma, A.H., 2015. Developmental immunotoxicity testing of 4-methyl anisole. Regul. Toxicol. Pharmacol. 72, 379–85.
276. Törnqvist E, Annas A, Granath B, Jalkesten E, Cotgreave I, Öberg M. Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. PLoS One. 2014 Jul 23;9(7):e101638. doi: 10.1371/journal.pone.0101638. eCollection 2014.
277. Tran TC, Sneed B, Haider J, Blavo D, White A, Aiyejorun T, Baranowski TC, Rubinstein AL, Doan TN, Dingledine R, Sandberg EM. 2007. Automated, quantitative screening assay for antiangiogenic compounds using transgenic zebrafish. Cancer Res 67: 11386-11392.
278. Truong L, Harper SL, Tanguay RL. Evaluation of embryotoxicity using the zebrafish model. Methods in molecular biology. 2011. 691:271-9.
279. Tyl RW, Crofton K, Moretto Al, Moser V, Sheets LP, Sobotka JJ (2008) Identification and interpretation of developmental neurotoxicity effects. A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. Neurotoxicol Teratol 30:349-381
280. Ujházy E, Mach M, Navarová J, Brucknerová I, Dubovický M. Teratology - past, present and future. Interdiscip Toxicol. 2012 Dec;5(4):163-8. doi: 10.2478/v10102-012-0027-0.
281. US EPA, 2013. Part 158 Toxicology Data Requirements: Guidance for Neurotoxicity Battery, Subchronic Inhalation, Subchronic Dermal and Immunotoxicity Studies. Office of Pesticide Programs, U.S. Environmental Protection Agency.
282. Van den Bulck K, Hill A, Mesens N, Diekman H, de Schaepdrijver L, Lammens L. Zebrafish developmental toxicity assay: A fishy solution to reproductive toxicity screening, or just a red herring? 2011. Reproductive Toxicology. 32:213-9.
283. van Gestel CA, Otermann K, Canton JH. 1985. Relation between water solubility, octanol/water partition coefficients, and bioconcentration of organic chemicals in fish: a review. Regul Toxicol Pharmacol 5: 422-431.
284. Van Vliet E. Current standing and future prospects for the technologies proposed to transform toxicity testing in the 21st century. ALTEX. 28(1):17-44 (2011).
285. Vargesson N. Thalidomide-induced teratogenesis: history and mechanisms.Birth Defects Res C Embryo Today. 2015 Jun;105(2):140-56. doi: 10.1002/bdrc.21096. Epub 2015 Jun 4.
286. Verbueken E, Alsop D, Saad MA, Pype C, Van Peer EM, Casteleyn CR, Van Ginneken CJ, Wilson J, Van Cruchten SJ. In Vitro Biotransformation of Two Human CYP3A Probe Substrates and Their Inhibition during Early Zebrafish Development. Int J Mol Sci. 2017 Jan 22;18(1). pii: E217. doi: 10.3390/ijms18010217.
287. Verhaeghe T, Dillen L, Stieltjes H, Zwart L, Feyen B, Diels L, Vroman A, Timmerman P. (2017). The application of capillary microsampling in GLP toxicology studies. Bioanalysis, 10.4155/bio-2016-0297.
288. Verthelyi, D., 2006. Female's heightened immune status: estrogen, T cells, and inducible nitric oxide synthase in the balance. Endocrinology 147, 659–661.
289. Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, Ottinger MA, Vergauwen L, Whelan M. Adverse outcome pathway (AOP) development I: strategies and principles. Toxicol Sci. Dec;142(2):312-20. doi: 10.1093/toxsci/kfu199 (2014).

290. Vogel, R., Seidle, T., Speilmann, H., 2010. A modular one-generation reproduction study as a flexible testing system for regulatory safety assessment. Reprod. Toxicol. 29, 242–245.
291. Waye, A., Trudeau, V.L., 2011. Neuroendocrine disruption: more than hormones are upset. J. Toxicol. Environ. Health B 14, 270–291.
292. Weigt S, Huebler N, Strecker R, Braunbeck T, Broschard TH. Zebrafish (Danio rerio) embryos as a model for testing proteratogens. Toxicology. 2011. 281:25-36.
293. Westerfield M. The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Brachydanio rerio). 4th ed. Eugene: University of Oregon Press; 2000.
294. WHO, 2012. Guidance for Immunotoxicity Risk Assessment for Chemicals. Harmonization Project Document No. 10. World Health Organization, Geneva.
295. Wier P. 2006. Use of toxicokinetics in developmental and reproductive toxicology. In: Hood RD, editor. Handbook of developmental toxicology. New York: CRL Press, Inc. p 571–597.
296. Wilson J. 1977. Current status of teratology: general principles and mechanisms derived from animal studies. In: Wilson J, Fraser F, editors. Handbook of teratology. 4 vols. New York: Plenum Press. pp. 47-74.
297. Wilson J.G. - Embryological Considerations in Teratology. In "Teratology: Principles and Techniques"; Wilson, J.G. and Warkany, J., Eds.; The University of Chicago Press: Chicago, IL, 251-277 (1965).
298. Wise LD, Buschmann J, Feuston MH, Fisher JE, Hew KW, Hoberman AM, Lerman SA, Ooshima Y, Stump DG. Embryo-fetal developmental toxicity study design for pharmaceuticals. Birth Defects Research (Part B) 86:418-428, 2009.
299. Woo, D.C., Hoar, R.M. Apparent hydronephrosis as a normal aspect of renal development in late gestation of rats: the effect of methyl salicylate. Teratology, 6, 191-196 (1972).
300. Yoon M, Barton HA (2008) Predicting maternal rat exposures: how different are they? Toxicol Sci 102(1):15-32. doi:10.1093/toxsci/kfm286
301. Zhang C, Ball J, Panzica-Kelly J, Augustine-Rauch K. In vitro developmental toxicology screens: a report on the progress of the methodology and future applications. Chemical Research in Toxicology, 29, 534-544 (2016).
302. Zmarowski A, Beekhuijzen M, Lensen J, Emmen H (2012) Differential performance of Wistar Han and Sprague Dawley rats in behavioral tests: Differences in baseline behavior and reactivity to positive control agents. Reprod Toxicol 34:192–203. doi: 10.1016/j.reprotox.2012.05.091

Appendix 3

Dankwoord



I would like to thank

- ✓ Aldert, for being an inspirational mentor.
- ✓ My DART team and Charles River colleagues, for your hard work, dedication and wonderful collaboration.
- ✓ The ETS, for your great friendships.
- ✓ My friends and family, for always being there when I needed you.
- ✓ Papa, mama, Sabine and Jos, for giving me a great childhood and for your continuous support.
- ✓ Bas and Roos, for just being wonderful. Michiel for being the love of my life 🧡

Appendix 4

List of publications

LIST OF PUBLICATIONS

PEER-REVIEWED PUBLICATIONS

Steven Cassara, **Manon Beekhuijzen**, Bruce Beyer, Robert Chapin, Martina Dorau, Alan Hoberman, Eckart Krupp, Isabelle LeConte, Don Stedman, Christine Stethem, Daphne van den Oetelaar, Belen Tornesi. A multi-institutional study benchmarking the zebrafish developmental assay for prediction of embryotoxic plasma concentrations from rat embryo-fetal development studies. Under review.

Manon Beekhuijzen, Steffen Schneider, Narinder Barraclough, Nina Hallmark, Alan Hoberman, Sheri Lordi, Mary Moxon, Deborah Perks, Aldert H. Piersma, Susan L. Makris. The Urgency for Optimization and Harmonization of Thyroid Hormone Analyses and Their Interpretation in Developmental and Reproductive Toxicology Studies. Reproductive Toxicology 2018, accepted.

George Daston, Aldert Piersma, Leonello Attias, **Manon Beekhuijzen**, Connie Chen, Jennifer Foreman, Nina Hallmark, Isabelle Leconte. Best Practices for Developmental Toxicity Assessment for Classification and Labeling. Reproductive Toxicology 2018, accepted.

George Daston, **Manon Beekhuijzen**. Is Omphalocele a Non-Specific Malformation in New Zealand White Rabbits? Reproductive Toxicology 78 (2018) 29-39.

Manon Beekhuijzen, Yaomi de Lange, Ankie Lambregts, Birgit Peter, Mira Wenker, Harry Emmen. Satellite rats are redundant in embryo-fetal development studies. Reproductive Toxicology, 2017 Sep;72:122-128

Manon Beekhuijzen. The era of 3Rs implementation in Developmental and Reproductive Toxicity (DART) Testing: current overview and future perspectives. Reproductive Toxicology, 2017 Sep;72:86-96.

Nigel P. Moore, **Manon Beekhuijzen**, Peter J. Boogaard, Jennifer E. Foreman, Colin M. North, Christine Palermo, Steffen Schneider, Volker Strauss, Bennard van Ravenzwaay, Alan Poole. Guidance on the selection of cohorts for the extended one-generation reproduction toxicity study (OECD test guideline 443). Regulatory Toxicology and Pharmacology 80 (2016) 32-40

Manon Beekhuijzen, Francois van Otterdijk, Willemien Wieland, Miranda van Tuyl, Robert Pels Rijcken, Birgit Peter, Harry Emmen. Update of OECD DART guidelines with endocrine disruptor relevant endpoints: Practical considerations. Reproductive Toxicology 64 (2016) 64–71

Xiaoyu Niu, **Manon Beekhuijzen**, Willem Schoonen, Harry Emmen, and Mira Wenker. Effects of Capillary Microsampling on Toxicological Endpoints in Juvenile Rats. Toxicological Sciences, 2016, 1–9

Manon Beekhuijzen, Helma Barentsen, Edward Marsden, Amy Zmarowski, Michel Aujoulat, Catherine Picut, Eddie Sloter. Implementing the extended one-generation reproductive toxicity study (EOGRTS): important points to consider. Critical Reviews in Toxicology, 2016

Coco de Koning, **Manon Beekhuijzen**, Marysia Tobor-Kapłon, Selinda de Vries-Buitenweg, Dick Schoutsen, Nico Leeijen, Beppy van de Waart, and Harry Emmen. Visualizing Compound Distribution during Zebrafish Embryo Development: The Effects of Lipophilicity and DMSO. Birth Defects Research (Part B) 00:1–20 (2015)

Michiel Alexander de Raaf, **Manon Beekhuijzen**, Christophe Guignabert, Anton Vonk Noordegraaf, Harm Jan Bogaard. Endothelin-1 receptor antagonists in fetal development and pulmonary arterial hypertension. Reproductive Toxicology (2015)

Manon Beekhuijzen, Coco de Koning, Maria-Eugenia Flores-Guillénc, Selinda de Vries-Buitenweg, Marysia Tobor-Kaplon, Beppy van de Waart, Harry Emmen. From cutting edge to guideline: A first step in harmonization of the zebrafish embryotoxicity test (ZET) by describing the most optimal test conditions and morphology scoring system. Reproductive Toxicology 56 (2015) 64–76

Manon Beekhuijzen, Michiel Alexander de Raaf, Amy Zmarowski, Francois van Otterdijk, Birgit Peter, Harry Emmen. The underestimated value of OECD 421 and 422 repro screening studies: putting it in the right perspective. Reproductive Toxicology 2014

Amy Zmarowski, **Manon Beekhuijzen**, Joost Lensen, Harry Emmen. Differential performance of Wistar Han and Sprague Dawley rats in behavioral tests: Differences in baseline behavior and reactivity to positive control agents. Reproductive Toxicology 2012

E. Rorije, A. Muller, **M. E.W. Beekhuijzen**, U. Hass, B. Heinrich-Hirsch M. Paparella, E. Schenk, B. Ulbrich, B.C. Hakkert, A.H. Piersma. Impact of 2nd generation mating and offspring on C&L, On the impact of second generation mating and offspring in multi-generation reproductive toxicity studies on Classification and Labelling of substances in Europe. Regulatory Toxicology and Pharmacology, 18 August 2011

A.H. Piersma, E. Rorije, **M.E.W. Beekhuijzen**, R. Cooper, D.J. Dix, B. Heinrich-Hirsch, M.T. Martin, E. Mendez, A. Muller, M. Paparella, D. Ramsingh, E. Reaves, P. Ridgway, E. Schenk, L. Stachiw, B. Ulbrich, B.C. Hakkert. Combined retrospective analysis of 498 rat multi-generation reproductive toxicity studies: On the impact of parameters related to F1 mating and F2 offspring. Reproductive Toxicology; 31 (4): 392-401, 2011

Manon Beekhuijzen, Amy Zmarowski, Harry Emmen, Wilbert Frieling. To mate or not to mate: A retrospective analysis of two-generation studies for evaluation of criteria to trigger additional mating in the extended one-generation design. Reproductive Toxicology 28, 203–208, 2009

E.J.H. Mulder, P.G. Robles de Medina, **M.E.W. Beekhuijzen**, D.E. Wijnberger, G.H.A. Visser. Fetal stimulation and activity state. The Lancet, Vol. 357, February 10, 478-479, 2001

M.E.W. Beekhuijzen, A. Verhoef, R. Klaasen, C.J.M. Rompelberg, A.H. Piersma. Comparison of in vitro and in vivo Developmental Toxicity and Pharmacokinetecs of Phenytoin in the rat. Pharmacology & Toxicology 87: 96-102, 2000

ORAL PRESENTATIONS

Manon Beekhuijzen. Developmental and Reproductive toxicity testing - the next generation. European Teratology Society (ETS), 4-7 September 2017, Budapest, Hungary

Manon Beekhuijzen. An introduction to the Extended One-Generation Reproductive Toxicity Study (EOGRTS). EDC expert workshop, 30 May-01 June 2017, Brussels, Belgium

Manon Beekhuijzen. OECD DART guidelines update: addition of endocrine disruptor relevant endpoints. Industrial Reproductive Toxicology Discussion Group (IRDG), 09-10 May 2017, Birmingham, UK

Manon Beekhuijzen. OECD DART guidelines update: addition of endocrine disruptor relevant endpoints. European Teratology Society (ETS), 11-14 September 2016, Dublin, Ireland

Manon Beekhuijzen. Visualizing compound distribution in zebrafish embryos: the effect of lipophilicity. Society of Toxicology (SOT), 22-26 March 2015, San Diego, California

Manon Beekhuijzen. Introduction of adverse outcome pathways (AOPs). European Teratology Society (ETS), 01-04 September 2014, Hamburg, Germany

Manon Beekhuijzen. Education course: Introduction into biopharmaceuticals and DART. European Teratology Society (ETS), 08-11 September 2013, Stresa, Italy

Manon Beekhuijzen. Different approaches in DART testing driven by biopharmaceutical subclasses. BelTox Second Workshop on Biopharmaceuticals, 22 November 2012, Braine-l'Alleud, Belgium

Manon Beekhuijzen. Implementing the Extended One-Generation Reproductive Toxicity Study (EOGRTS). Important points to consider. European Teratology Society (ETS), 02-05 September 2012, Linz, Austria

Manon Beekhuijzen. DART under REACH. Nederlandse Vereniging voor Toxicologie (NVT) sectie teratologie en reproductietoxicologie, 06 March 2012, Den Bosch, the Netherlands

Manon Beekhuijzen. Postnatal assessments in juvenile rabbits. J&J Juvenile Toxicology Symposium (Satellite meeting of ETS), 07-08 September 2011, Gent, Belgium

Manon Beekhuijzen. To mate or not to mate: A retrospective analysis of fourteen two-generation studies for evaluation of the added value of a second generation. An update. European Teratology Society (ETS), 05-08 September 2010, Barcelona, Spain

POSTER PRESENTATIONS

Steffi Bressers, Marysia Tobor-Kaplon, **Manon Beekhuijzen**. Optimization of the use of Caenorhabditis elegans (C. elegans) in reproductive toxicology testing: describing the most optimal test conditions. Nederlandse Vereniging voor Toxicologie (NVT), 30-31 mei 2018, Hilversum, the Netherlands

Kirsten Hartman-Van Dycke, Francois van Otterdijk, Chantal Beerens, Harrie van der Eerden, Harry Emmen, **Manon Beekhuijzen**. Optimization of semi-occlusive dermal exposure method in pregnant and lactating rats. European Societies of Toxicology (Eurotox), 10-13 September 2017, Bratislava, Slovakia

Kirsten Hartman-Van Dycke, Francois van Otterdijk, Chantal Beerens, Harrie van der Eerden, Harry Emmen, **Manon Beekhuijzen**. Enhanced method of dermal exposure (semi-occlusive) in rats during gestation and lactation: Improved animal welfare and study outcome. European Teratology Society (ETS), 04-07 September 2017, Budapest, Hungary

Manon Beekhuijzen, Yaomi de Lange, Michel Blocks, Mira Wenker, Harry Emmen. Are Satellite Rats Needed for Toxicokinetic Assessment in Embryo-Fetal Development Studies? Society of Toxicology (SOT), March 12-16, 2017, Baltimore, MD

Manon Beekhuijzen, Linda Allais, Sarah Davis, Marilene Leroy, Pragati Coder, Kirsten Hartman-Van Dycke. Evaluating the Use of the Second Species in Embryo-Fetal or Prenatal Developmental Studies. European Teratology Society (ETS), 10-14 September 2016, Dublin, Ireland

Xiaoyu Niu, Mira Wenker, **Manon Beekhuijzen**, Harry Emmen. Effect of capillary microsampling on toxicological endpoints in juvenile rats at PND4, 10 and 17. European Societies of Toxicology (Eurotox), 13-16 September 2015, Porto, Portugal

Manon Beekhuijzen, Edward Marsden, Eddie Sloter. The Extended One-Generation Reproductive Toxicity Study (EOGRTS): An Update on Current European Legislation and Considerations for Implementation. Teratology Society (TS), 27 June – 01 July 2015, Montreal, Quebec, Canada

Mira Wenker, Xiaoyu Niu, **Manon Beekhuijzen**, Harry Emmen. Microsampling in Juvenile Rats on PND4, 10 and 17. Society of Toxicology (SOT), 22-26 March 2015, San Diego, California

Coco de Koning, **Manon Beekhuijzen**, Marysia Tobor-Kapłon, Beppy van de Waart, Harry Emmen. Visualizing compound distribution during zebrafish embryo development. Society of Environmental Toxicology and Chemistry (SETAC), 03-07 May 2015, Barcelona, Spain

Manon Beekhuijzen, Coco de Koning, Selinda de Vries-Buitenweg, Marysia ToborKapłon, Dick Schoutsen, Nico Leeijen, Beppy van de Waart, Harry Emmen. Zebrafish embryotoxicity test (ZET): the importance of internal exposure analyses. European Teratology Society (ETS), 01-04 September 2014, Hamburg, Germany

Manon Beekhuijzen, Maru Flores-Guillén, Harry Emmen. Selection of the most optimal test conditions for a high predictability zebrafish embryotoxicity test. Belgian Society of Toxicology and Ecotoxicology (BelTox), 06 December 2013, Brussels, Belgium

Amy Zmarowski, **Manon Beekhuijzen**, Harry Emmen. Differential Performance of Wistar Han and Sprague Dawley Rats in Behavioral Tests: Differences in Baseline Behavior and Reactivity to Positive Control Agents. International Neurotoxicology Association (INA), 2013, Amsterdam, the Netherlands

Amy Zmarowski, **Manon Beekhuijzen**, Marleen Teunissen, Harry Emmen. Sensorimotor function in juvenile and adult Sprague Dawley and Wistar Han rats. European Societies of Toxicology (Eurotox), 17-20 June 2012, Stockholm, Sweden

Amy Zmarowski, **Manon Beekhuijzen**, Marleen Teunissen, Alex Eapen, Harry Emmen. Characterization of motor activity over the lifespan in Sprague Dawley and Wistar Han rats. Society of Toxicology (SOT), 11-15 March 2012, San Francisco, California

Manon Beekhuijzen, Maartje Pennings, Birgit Peter, Bianca van Rozendaal, Harry Emmen. Development of postnatal assessments in juvenile rabbits. European Teratology Society (ETS), 04-07 September 2011, Gent, Belgium

Amy Zmarowski, **Manon Beekhuijzen**, Marleen Teunissen, Harry Emmen. Differential effects of Methylazoxymethanol and MK-801 administration on learning and memory impairment in Sprague Dawley and Wistar Han rats. European Societies of Toxicology (Eurotox), 28-31 August 2011, Paris, France

A.H. Piersma, E. Rorije, **M.E.W. Beekhuijzen**, R. Cooper, D.J. Dix, B. Heinrich-Hirsch, M.T. Martin, E. Mendez, A. Muller, M. Paparella, D. Ramsingh, E. Reaves, P. Ridgway, E. Schenk, L. Stachiw, B. Ulbrich, B.C. Hakkert. Combined retrospective analysis of 498 rat multi-generation reproductive

toxicity studies: On the impact of parameters related to F1 mating and F2 offspring. Teratology Society (TS), 25-29 June 2011, Coronado, California

Manon Beekhuijzen, Amy Zmarowski, Bianca van Rozendaal, Michiel de Raaf, Harry Emmen. Bi-Directional Changes in motor activity and startle response in juvenile and adult rats. European Societies of Toxicology (Eurotox), 19-23 July 2010, Barcelona, Spain

Amy Zmarowski, **Manon Beekhuijzen**, Michiel de Raaf, Harry Emmen, Wilbert Frieling. A Validation study to assess the effect of amphetamine and the typical antipsychotic chlorpromazine on startle response and locomotor activity in juvenile and adult rats. Nederlandse Vereniging voor Toxicologie (NVT), 2009, the Netherlands

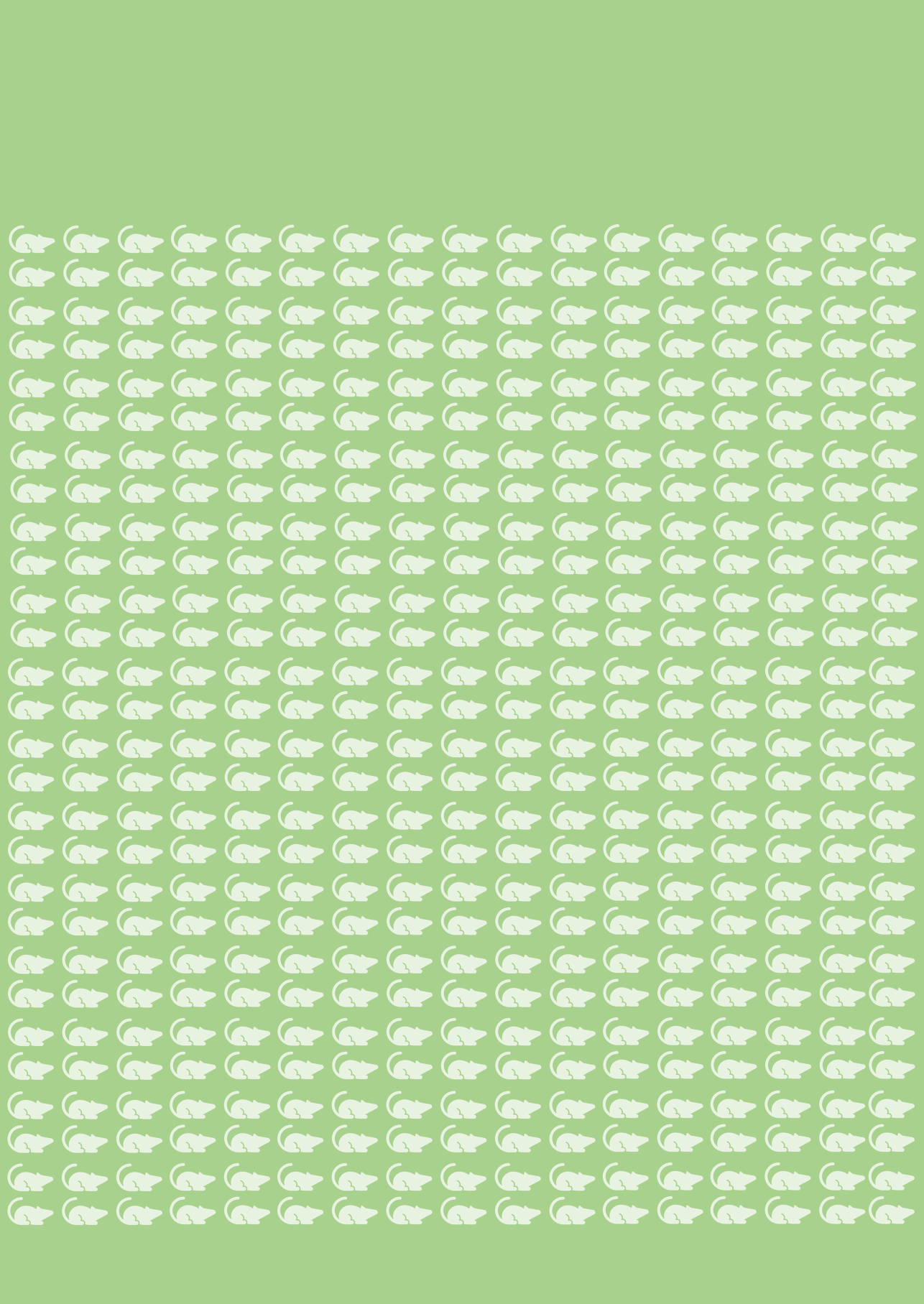
Wilbert Frieling, Liesbeth Heijink, Sheila Tesh, Beverly Leggett, **Manon Beekhuijzen**. Embryofoetal developmental toxicity of propylene glycol in NZW rabbits. European Teratology Society (ETS), September 2000, Bologna, Italy

Appendix 5

Curriculum Vitae

CURRICULUM VITAE

Manon Beekhuijzen studied medical biology (1995-1999) at the University of Utrecht of which her major internship was performed at the National Institute for Public Health and the Environment (RIVM) under supervision of Aldert Piersma. She carried out an extra internship at Cornell University in Ithaca, New York, USA under supervision of Peter Nathanielsz.

In 1999 she joined Charles River (then named NOTOX), a Contract Research Organization (CRO) in the Netherlands, as Study Director for developmental and reproduction toxicology (DART) studies. She was responsible for a major part in setting up all different type of DART and juvenile studies at this site, and she supervised eight students. From 2014 onwards, she was promoted as section head DART and is currently managing a group of nine DART study directors.

She is registered as European Toxicologist (ERT), current president of the DART subgroup of the Dutch Society of Toxicology (NVT), and was president of the European Teratology Society (ETS) from 2015 to 2016.