Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

13-Apr-2012

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

Cancels & replaces the same document of 03 April 2012

GUIDANCE DOCUMENT 116 ON THE CONDUCT AND DESIGN OF CHRONIC TOXICITY AND CARCINOGENICITY STUDIES, SUPPORTING TEST GUIDELINES 451, 452 AND 453
2ND EDITION

Series on Testing and Assessment

No. 116

JT03319769

Complete document available on OLIS in its original format
This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

ENV/JM/MONO(2011)47
Unclassified
English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 116

GUIDANCE DOCUMENT 116 ON THE CONDUCT AND DESIGN OF CHRONIC TOXICITY AND CARCINOGENICITY STUDIES, SUPPORTING TEST GUIDELINES 451, 452 AND 453 – 2ND EDITION

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2012

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (1993; reformatted 1995, revised 2006)*

No. 2, *Detailed Review Paper on Biodegradability Testing (1995)*

No. 3, *Guidance Document for Aquatic Effects Assessment (1995)*

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment (1995)*

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing (1996)*

No. 6, *Report of the Final Ring-test of the Daphnia magna Reproduction Test (1997)*

No. 7, *Guidance Document on Direct Phototransformation of Chemicals in Water (1997)*

No. 8, *Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)*

No. 9, *Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides during Agricultural Application (1997)*

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (2004)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (2007)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (2005)*

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances Which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (in preparation)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures Which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (2003)*

No. 43, *Guidance Document on Mammalian Reproductive Toxicity Testing and Assessment (2008)*

No. 44, *Description of Selected Key Generic Terms Used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

No. 46, *Detailed Review Paper on Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances (2004)*

No. 47, *Detailed Review Paper on Fish Screening Assays for the Detection of Endocrine Active Substances (2004)*

No. 48, *New Chemical Assessment Comparisons and Implications for Work Sharing (2004)*

No. 49, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs (2004)*

No. 50, *Report of the OECD/IPCS Workshop on Toxicogenomics (2005)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 52, *Comparison of emission estimation methods used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case study of pulp and paper and textile sectors (2006)*

No. 53, *Guidance Document on Simulated Freshwater Lentic Field Tests (Outdoor Microcosms and Mesocosms) (2006)*

No. 54, *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (2006)*

No. 55, *Detailed Review Paper on Aquatic Arthropods in Life Cycle Toxicity Tests with an Emphasis on Developmental, Reproductive and Endocrine Disruptive Effects (2006)*

No. 56, *Guidance Document on the Breakdown of Organic Matter in Litter Bags (2006)*

No. 57, *Detailed Review Paper on Thyroid Hormone Disruption Assays (2006)*

No. 58, *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals (2006)*

No. 59, *Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-Day Oral Toxicity Study in Laboratory Rats (2006)*

No. 60, *Report of the Initial Work Towards the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1A) (2006)*

No. 61, *Report of the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1B) (2006)*

No. 62, *Final OECD Report of the Initial Work Towards the Validation of the Rat Hershberger Assay: Phase-1, Androgenic Response to Testosterone Propionate, and Anti-Androgenic Effects of Flutamide (2006)*

No. 63, *Guidance Document on the Definition of Residue (2006)*

No. 64, *Guidance Document on Overview of Residue Chemistry Studies (2006)*

No. 65, *OECD Report of the Initial Work Towards the Validation of the Rodent Uterotrophic Assay - Phase 1 (2006)*

No. 66, *OECD Report of the Validation of the Rodent Uterotrophic Bioassay: Phase 2. Testing of Potent and Weak Oestrogen Agonists by Multiple Laboratories (2006)*

No. 67, *Additional data supporting the Test Guideline on the Uterotrophic Bioassay in rodents (2007)*

No. 68, *Summary Report of the Uterotrophic Bioassay Peer Review Panel, including Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the follow up of this report (2006)*

No. 69, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models (2007)*

No. 70, *Report on the Preparation of GHS Implementation by the OECD Countries (2007)*

No. 71, *Guidance Document on the Uterotrophic Bioassay - Procedure to Test for Antioestrogenicity (2007)*

No. 72, *Guidance Document on Pesticide Residue Analytical Methods (2007)*

No. 73, *Report of the Validation of the Rat Hershberger Assay: Phase 3: Coded Testing of Androgen Agonists, Androgen*

Antagonists and Negative Reference Chemicals by Multiple Laboratories. Surgical Castrate Model Protocol (2007)

No. 74, *Detailed Review Paper for Avian Two-generation Toxicity Testing (2007)*

No. 75, *Guidance Document on the Honey Bee (Apis Mellifera L.) Brood test Under Semi-field Conditions (2007)*

No. 76, *Final Report of the Validation of the Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances: Phase 1 - Optimisation of the Test Protocol (2007)*

No. 77, *Final Report of the Validation of the Amphibian Metamorphosis Assay: Phase 2 - Multi-chemical Interlaboratory Study (2007)*

No. 78, *Final Report of the Validation of the 21-day Fish Screening Assay for the Detection of Endocrine Active Substances. Phase 2: Testing Negative Substances (2007)*

No. 79, *Validation Report of the Full Life-cycle Test with the Harpacticoid Copepods Nitocra Spinipes and Amphiascus Tenuiremis and the Calanoid Copepod Acartia Tonsa - Phase 1 (2007)*

No. 80, *Guidance on Grouping of Chemicals (2007)*

No. 81, *Summary Report of the Validation Peer Review for the Updated Test Guideline 407, and Agreement of the Working Group of National Coordinators of the Test Guidelines Programme on the follow-up of this report (2007)*

No. 82, *Guidance Document on Amphibian Thyroid Histology (2007)*

No. 83, *Summary Report of the Peer Review Panel on the Stably Transfected Transcriptional Activation Assay for Detecting Estrogenic Activity of Chemicals, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 84, *Report on the Workshop on the Application of the GHS Classification Criteria to HPV Chemicals, 5-6 July Bern Switzerland (2007)*

No. 85, *Report of the Validation Peer Review for the Hershberger Bioassay, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 86, *Report of the OECD Validation of the Rodent Hershberger Bioassay: Phase 2: Testing of Androgen Agonists, Androgen Antagonists and a 5 α-Reductase Inhibitor in Dose Response Studies by Multiple Laboratories (2008)*

No. 87, *Report of the Ring Test and Statistical Analysis of Performance of the Guidance on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (Transformation/ Dissolution Protocol) (2008)*

No. 88, *Workshop on Integrated Approaches to Testing and Assessment (2008)*

No. 89, *Retrospective Performance Assessment of the Test Guideline 426 on Developmental Neurotoxicity (2008)*

No.90, *Background Review Document on the Rodent Hershberger Bioassay (2008)*

No. 91, *Report of the Validation of the Amphibian Metamorphosis Assay (Phase 3) (2008)*

No. 92, *Report of the Validation Peer Review for the Amphibian Metamorphosis Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-Up of this Report (2008)*

No. 93, *Report of the Validation of an Enhancement of OECD TG 211: Daphnia Magna Reproduction Test (2008)*

No. 94, *Report of the Validation Peer Review for the 21-Day Fish Endocrine Screening Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2008)*

No. 95, *Detailed Review Paper on Fish Life-Cycle Tests (2008)*

No.96, *Guidance Document on Magnitude of Pesticide Residues in Processed Commodities (2008)*

No.97, *Detailed Review Paper on the use of Metabolising Systems for In Vitro Testing of Endocrine Disruptors (2008)*

No. 98, *Considerations Regarding Applicability of the Guidance on Transformation/Dissolution of Metals Compounds in Aqueous Media (Transformation/Dissolution Protocol) (2008)*

No. 99, *Comparison between OECD Test Guidelines and ISO Standards in the Areas of Ecotoxicology and Health Effects (2008)*

No. 100, *Report of the Second Survey on Available Omics Tools (2009)*

No. 101, *Report of the Workshop on Structural Alerts for the OECD (Q)SAR Application Toolbox, 15-16 May 2008, Utrecht, the Netherlands (2009)*

No. 102, *Guidance Document for using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals (2009)*

No. 103, *Detailed Review Paper on Transgenic Rodent Mutation Assays (2009)*

No. 104, *Performance Assessment: Comparsion of 403 and CxT Protocols via Simulation and for Selected Real Data Sets (2009)*

No. 105, *Report on Biostatistical Performance Assessment of the draft TG 436 Acute Toxic Class Testing Method for Acute Inhalation Toxicity (2009)*

No. 106, *Guidance Document for Histologic Evaluation of Endocrine and Reproductive Test in Rodents (2009)*

No. 107, *Preservative treated wood to the environment for wood held in storage after treatment and for wooden commodities that are not cover and are not in contact with ground. (2009)*

No. 108, *Report of the validation of the Hershberger Bioassay (weanling model) (2009)*

No. 109, *Literature review on the 21-Day Fish Assay and the Fish Short-Term Reproduction Assay (2009)*

No. 110, *Report of the validation peer review for the weanling Hershberger Bioassay and agreement of the working of national coordinators of the test guidelines programme on the follow-up of this report (2009)*

No. 111, *Report of the Expert Consultation to Evaluate an Estrogen Receptor Binding Affinity Model for Hazard Identification (2009)*

No. 112, *The 2007 OECD List of High Production Volume Chemicals (2009)*

No. 113, *Report of The Focus Session on Current and Forthcoming Approaches for Chemical Safety and Animal Welfare (2010)*

No. 114, *Performance Assessment of Different Cytotoxic and Cytostatic Measures for the In Vitro Micronucleus Test (MNVIT): Summary of results in the collaborative trial (2010)*

No. 115, *Guidance Document on the Weanling Hershberger Bioassay in Rats: A Short-term Screening Assay for (Anti) Androgenic Properties (2009)*

No. 116, *Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453 – 2nd Edition (2011)*

No. 117, *Guidance Document 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 (2011)*

No. 118, *Workshop Report on OECD Countries Activities Regarding Testing, Assessment and Management of Endocrine Disrupters Part I and Part II (2010)*

No. 119, *Classification and Labelling of chemicals according to the UN Globally Harmonized System: Outcome of the Analysis of Classification of Selected Chemicals listed in Annex III of the Rotterdam Convention (2010)*

No. 120, *Part 1: Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism-based Structural Alerts for the Identification of DNA Binding Chemicals*

No. 120, *Part 2: Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism-based Structural Alerts for the Identification of DNA Binding Chemicals*

No. 121, *Detailed review paper (DRP) on Molluscs life-cycle Toxicity Testing (2010)*

No. 122, *Guidance Document on the determination of the Toxicity of a Test Chemical to the Dung Beetle Aphodius Constans (2010)*

No. 123, *Guidance Document on the Diagnosis of Endocrine-related Histopathology in Fish Gonads (2010)*

No. 124, *Guidance for the Derivation of an Acute Reference Dose (2010)*

No. 125, *Guidance Document on Histopathology for Inhalation Toxicity Studies, Supporting TG 412 (Subacute Inhalation Toxicity: 28-Day) and TG 413 (Subchronic Inhalation Toxicity: 90-Day) (2010)*

No. 126, S*hort Guidance on the Threshold approach for Acute Fish Toxicity (2010)*

No. 127, *Peer review report of the validation of the 21-day androgenised female stickleback screening assay (2010)*

No. 128, *Validation Report of the 21-day Androgenised Female Stickleback Screening Assay (2010)*

No. 129, *Guidance Document on using Cytotoxicity Tests to Estimate Starting Doses for Acute Oral Systemic Toxicity Tests(2010)*

No. 131, *Report of the Test Method Validation of Avian Acute Oral Toxicity Test (OECD test guideline 223) (2010)*

No. 132, *Report of the Multi-Laboratory Validation of the H295R Steroidogenesis Assay to Identify Modulators (2010)*

No.133, *Peer Review Report for the H295R Cell-Based Assay for Steroidogenesis (2010)*

No.134, *Report of the Validation of a Soil Bioaccumulation Test with Terrestrial Oligochaetes by an International ring test (2010)*

No.135, *Detailed Review Paper on Environmental Endocrine Disruptor Screening: The use of Estrogen and Androgen Receptor Binding and Transactivation Assays in Fish (2010)*

No. 136, *Validation Report of t he Chironomid Full Life-Cycle Toxicity Test (2010)*

No. 137, *Explanatory Background Document to the OECD Test Guideline On In Vitro Skin Irritation Testing (2010)*

No. 138, *Report of the Workshop on Using Mechanistic Information in Forming Chemical Categories (2011)*

No. 139, *Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism Based Structural Alerts for the Identification of Protein-binding Chemicals (2011)*

No. 140, *Report of the WHO/OECD/ILSI (Hesi) Workshop on Risk Assessment of Combined Exposures to Multiple Chemicals (2011)*

No. 141, *Report of the Phase 1 of the Validation of the Fish Sexual Development Test for the Detection of Endocrine Active Substances (2011)*

No. 142, *Report of the Phase 2 of the Validation of the Fish Sexual Development Test for the Detection of Endocrine Active Substances (2011)*

No. 143, *Peer Review Report for the Validation of the Fish Sexual Development Test and Agreement of the Working Group of National Co-ordinators of the Test Guideline Programme on the Follow-up of the Peer Review (2011)*

No. 144, *Validation Report for the Acute Chironomid Assay (2011)*

No. 145, *Transgenic Rodent Somatic and Germ Cell Gene Mutation Assay: Retrospective Performance Assessment (2011)*

No. 146, *Syrian Hamster Embryonic (SHE) Cell PH 6.7 Cell Transformation Assay Prevalidation Study Report (2011)*

No. 147, *Syrian Hamster Embryonic (SHE) Cell PH 7.0 Cell Transformation Assay Prevalidation Study Report (2011)*

No. 148, *Guidance Document on the Androngenised Female Stickleback Screen (2011)*

No. 149, *Validation Report of the Balb/c 3T3 Cell Transformation Assay (2011)*

No. 150, *Guidance Document on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption (2011)*

No. 152, *Case Study: Assessment of an Extended Chemical Category, the Short-chain Methacrylates, Targeted on Bioaccumulation (2011)*

No. 153, *Guidance Document for the Derivation of an Acute Reference Concentration (Arfc) (2011)*

No. 154, *Draft Validation Report: Part 1 – Validation of Efficacy Methods for Antimicrobials used on Hard Surfaces (2011)*

No. 154, *Draft Validation Report: Part 2 – Validation of Efficacy Methods for Antimicrobials used on Hard Surfaces (2011)*

No. 155, *Peer Review for the Validation of the Modified Skin Irritation Test Method using LabyCyte EPI-MODEL24; additional studies; and agreement of the Working Group of National Coordinators on the follow-up to the Peer Review*

No. 156, *Guidance Notes on Dermal Absorption* (2011)

No. 157, *Validation Report Phase 1 for the Zebrafish Embryo Toxicity Test (2011)*

No. 158, *Report of Progress on the Interlaboratory Validation of the OECD Harpacticoid Copepod Development and Reproduction Test (2011)*

No. 159, *Validation Report for the Skin Irritation Test Method using Labcyte Epi-Model24 (2011)*

No. 160, *Guidance Document on the Bovine Corneal Opacity and Permeability (Bcop) and Isolated Chicken Eye (Ice) Test Methods: Collection of Tissues For Histological Evaluation And Collection Of Data On Non-Severe Irritants (2011)*

No. 161, *Peer Review Report for the Validation of the Stably Transfected Transcriptional Activation Assay for the Detection Androgenic and Anti-Androgenic Activity of Chemicals (2011)*

© OECD 2012

Applications for permission to reproduce or translate all or part of this material should be made to: Head of Publications Service, RIGHTS@oecd.org. OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France

ABOUT THE OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in ten different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides and Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/ehs/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/ehs/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document is the second edition of the Guidance Document (GD) 116 on the Design and Conduct of chronic toxicity and carcinogenicity studies, supporting Test Guidelines 451, 452 and 453 (carcinogenicity, chronic toxicity and combined chronic toxicity/carcinogenicity studies).

The proposal for developing this GD was approved by the WNT in 1997. At that time the project only included the development of a GD on dose selection. In 2008, the objective of the project was revised. The WNT agreed that the GD should be developed in parallel with the update of the Test Guidelines 451, 452 and 453 as a supporting GD for these Test Guidelines. Thus, although the WNT agreed that the section on dose selection should be developed as a priority, work also started to develop other areas of guidance.

The proposal for this Guidance Document was first discussed, with the draft updated Test Guidelines 451, 452 and 453, at a workshop held in Washington D.C. in 2008. The 1st edition including the general introduction (Chapter 1) and the section on dose selection (Section 3.1) was finalized at an expert meeting held in Paris on 7-8 October 2009 and published in June 2010. This 2nd edition, including the entire GD, was finalised at an expert meeting held in Paris in November 2010. WNT comments on two successive drafts were requested before the expert meeting and the entire draft GD was circulated for a third commenting round to the WNT in December 2010.

Comments from the WNT have been addressed and the second edition of the GD was approved by the WNT at its meeting held in April 2011. The Joint Meeting of the Chemicals Committee and the working Party on Chemicals, Pesticides and Biotechnology agreed to the declassification of the second edition of this document in Ocotber 2011.

This document is published under the responsibility of the Joint Meeting of the Chemicals Committee and the working Party on Chemicals, Pesticides and Biotechnology.

TABLE OF CONTENTS

ABOUT THE OECD 16
FOREWORD 18
1. GENERAL INTRODUCTION 22
1.1 Guiding principles and considerations 22
1.2 Scope of application of the guidance 24
1.3 Objectives of a chronic toxicity study 24
1.4 Objectives of a carcinogenicity study 24
1.5 Objectives of a combined chronic toxicity/carcinogenicity study 25
1.6 Consideration of testing strategies 25
1.7 Animal welfare considerations 27
2. MODE OF TOXICOLOGIC ACTION 33
2.1 Introduction 33
2.2 Background 33
2.3 "Mode of Action" Framework 34
2.4 Human Relevance Framework 37
2.5 Consideration of "Mode of Action" information to Optimise the Design of Long-term Studies 39
Appendix I. Examples of Animal Mode(s) of Action (MOA) Framework 44
Appendix II. Example of Human Relevance Framework 48
3. STUDY DESIGN 49
3.1 DOSE SELECTION 49
3.1.1 Introduction 49
3.1.2 Principles for Dose Selection 50
3.1.3 Integration of the objectives of a long-term bioassay 56
Appendix 1: ILSI Principles for Dose Selection in Chronic Rodent Bioassays (Rhomberg et al., 2007) 61
3.2 ROUTES OF EXPOSURE AND DOSE ADMINISTRATION CONSIDERATIONS 66
3.2.1 The oral route of exposure 66
3.2.2 The dermal route of exposure 69
3.2.3 The inhalation route of exposure 69

3.2.4 Other routes of exposure 71

3.3 CHOICE OF SPECIES AND STRAIN, NUMBERS AND SEX OF ANIMALS, STUDY DURATION, ALTERNATIVE *IN VIVO* MODELS 75

3.3.1 Testing in rodents 75

3.3.2 Testing in non-rodents, including considerations of numbers of animals to be used 78

3.3.3 Study duration 79

3.3.4 Alternative in vivo models for carcinogenicity testing, including testing in transgenic animals 80

3.4 TOXICOKINETICS 86

3.5 HOUSING, FEEDING, HANDLING OF ANIMALS AND EXPERIMENTAL PROCEDURES 91

3.5.1 Housing 91

3.5.2 Feeding 93

3.5.3 Handling, Health Surveillance and Experimental Procedures 95

3.6 INVESTIGATIONS (INCLUDING HISTOPATHOLOGICAL GUIDANCE) 99

3.6.1 Introduction 99

3.6.2 Ophthalmoscopy 99

3.6.3 Pathology and Histopathology 100

3.6.4 Supplementary Investigations 107

3.6.5 Avoiding bias 107

3.6.6 Data Archives 108

4. STATISTICAL AND DOSE RESPONSE ANALYSIS, INCLUDING BENCHMARK DOSE AND LINEAR EXTRAPOLATION, NOAELS AND NOELS, LOAELS AND LOELS 114

4.1 Preamble 114

4.2 Introduction 114

4.3 Objectives 115

4.4 Current statements on statistical methods 115

4.5 Study designs 116

4.6 Control groups and length of study 117

4.7 Purpose of statistical analyses 118

4.8 Types of data: Qualitative and Quantitative endpoints 119

4.9 Sample size and power considerations 120

4.10 Selection of Statistical Methods using Flowcharts 121

4.11 Description of the OECD flowchart 122

4.12 Intercurrent mortality 125

4.13 Cause of death (COD) and Context of observation (COO) 125

4.14 Time to tumour onset 126

4.15 Standard (simple) statistical analysis of qualitative data 127

4.16 Survival adjusted analyses 128

4.17 Tests of difference in survival 131

4.18 Assumptions for statistical analysis 132

4.19 One- or two-sided tests 133

4.20 Interpretation of statistical analyses 134

4.21 Use of control data and dual control groups 134

4.22 Historical control considerations 135

4.23 Dose-response modelling 136

4.24 Extrapolation to low doses 137

4.25 NOEL, NOAEL, LOEL, LOAEL approach 137

4.26 Benchmark dose approach 138

Appendix 1 Common statistical methods used in the analysis of data 143

5 DEFINITIONS/GLOSSARY 152

1. GENERAL INTRODUCTION

1.1 Guiding principles and considerations

1. Chronic toxicity and carcinogenicity studies are intended to identify toxic effects and potential health hazards following prolonged, repeated exposure. This type of study is usually required if humans are likely to be exposed to a substance over a significant portion of their life span. In the 1960s, long-term animal bioassays (chronic toxicity and carcinogenicity studies) began to be routinely used for hazard identification, to assess the qualitative potential of a chemical to cause chronic toxicity and cancer.

2. The objectives of the long-term bioassays have however expanded beyond hazard identification and are now focused primarily on hazard characterization for use in the assessment of risk for humans. In addition there has been increasing pressure for the long-term bioassay designs to consider financial constraints and societal desires to minimize the number of animals needed for scientific interpretation of results. There is a growing desire for long-term studies to provide data that cover a number of objectives including characterization of the nature of specific toxic responses, description of dose–response relationship, establishment of inflection points, and provision of insight into the roles of toxicokinetics and mechanisms of toxic action. In practice, it is likely that the bioassay design will be a compromise among a set of different purposes; to the extent that the ability to address one question is enhanced, the ability to address others may be diminished. For example, it may be necessary to achieve a balance between the power to detect toxicity and the ability to estimate the dose–response relationship of any observed effects. If information on carcinogenicity hazard identification is not available, this should be the main objective of the study.

3. The use of formal risk assessment procedures by government regulatory bodies began to emerge in the late 1970s and early 1980s bringing with it a strong interest in using data for quantitative as well as qualitative purposes. The need to gather data that allowed an understanding of the shape and slope of the dose-response curve focused attention on the number of doses in a bioassay and their spacing. Advances in knowledge of how chemicals perturbed or otherwise modulated biological processes in the development of tumours or other forms of toxicity provided bases for further improving the risk assessment process. Through meetings held primarily under the auspices of the International Programme on Chemical Safety (IPCS), a Mode of Action (MOA) framework was developed and refined (Sonnich-Mullin *et al*., 2001; Cohen *et al*., 2003; Meek *et al*., 2003; Holsapple *et al*., 2006; Boobis *et al*., 2006; EPA, 2005), as will be further developed in Chapter 2 of this guidance. The key purpose of this work was to introduce greater transparency into the process of assessing human relevance, and the goal was to use a broad array of relevant data to determine the predictive value of a bioassay tumour response to risk in humans.

4. The broadened range and complexity of scientific data used to evaluate chemical toxicity and carcinogenicity potential for humans highlighted the need to revise and update the following OECD Test Guidelines (TGs): TG 451 (Carcinogenicity Studies), TG 452 (Chronic Toxicity Studies), and TG 453–(Combined Chronic Toxicity/Carcinogenicity Studies), originally adopted in 1981. These TGs have therefore recently been revised in the light of scientific progress and the updating of related OECD Guidelines such as TG 408 (90-day oral toxicity study in rodents) and TG 407 (28-day oral toxicity study in rodents).

5. During the revision of the Test Guidelines, an emphasis was placed on providing guidance on factors that influence the selection of test doses, particularly for carcinogenicity studies. It was recognized that while general principles of dose selection should be contained in the Test Guidelines themselves, there was a need for additional guidance on these principles. The revision took into

account two publications by the International Life Sciences Institute (ILSI), "*Principles for the Selection of Doses in Chronic Rodent Bioassays"* (ILSI, 1997), and "*Issues in the Design and Interpretation of Chronic Toxicity and Carcinogenicity Studies in Rodents: Approaches to Dose Selection*" (Rhomberg *et al*., 2007). These reports provided theoretical and practical guidance on factors that influence dose selection in long-term bioassays.

6. A summary of the principles contained in these two publications, to underpin the texts on dose selection contained in the Test Guidelines, is provided in this guidance (Section 3.1, Appendix 1). During the development of this material, suggestions were made for additional guidance on specific aspects of study design in relation to core objectives of these studies, and how they might impact on other aspects of the study (e.g., designing for optimal collection of carcinogenicity data versus chronic toxicity data, design of studies for risk estimation rather than hazard assessment). It was generally agreed that the scope of the guidance should be wider than principles of dose selection, and should cover a number of key issues related to carcinogenicity and chronic toxicity testing.

7. This guidance therefore provides additional information on the conduct of studies performed using TG 451, 452 and TG 453. Its objective is to assist users of the TGs to select the most appropriate methodology to assess the chronic toxicity and carcinogenicity of a test chemical so that particular data requirements can be met while reducing animal usage if possible/appropriate.

8. The guidance is intended to foster a common approach among those carrying out chronic toxicity and carcinogenicity studies, and thereby contributes to the harmonisation activities undertaken by the OECD and other agencies, such as the WHO[1]. It should be consulted in addition to other guidance and requirements. It provides broad guidance on approaches to the execution of chronic toxicity and carcinogenicity studies. The text reflects current scientific understanding and standards. In time, the scientific community will gain a better understanding of the mechanisms of toxicity, and this may lead to changes in both methodology and interpretation of results, which should reflect scientific consensus at the time data are reviewed.

9. It should be noted that the basic principles for the conduct of chronic toxicity and carcinogenicity studies will differ, given that the endpoints are different. While the guidance provided in this document can be taken as generally applicable to the conduct of either a chronic toxicity or a carcinogenicity bioassay, or a combined chronic and carcinogenicity study, users of the guidance should be mindful of the primary objectives of the study. It should also be noted that much of the information provided in the guidance reflects general principles for conducting animal studies, e.g., information on route of exposure and dosing considerations, choice of species and strain, the importance of toxicokinetic studies, housing and feeding and animal welfare issues, and is not unique to chronic toxicity or carcinogenicity studies. The objective in restating this general information in this guidance is to provide as far as possible a stand-alone document, avoiding the need for the reader to make cross reference between several guidance documents covering different aspects of such studies.

10. Two other OECD documents also provide guidance on aspects of the conduct and interpretation of repeat-dose toxicity studies, chronic toxicity and carcinogenicity studies, Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (OECD, 2002a) and Guidance

---

[1] An example, already referred to, is the IPCS project *Harmonisation of Approaches to the Assessment of Risk from Exposure to Chemicals*, which has developed a *Conceptual Framework for Cancer Risk Assessment*. The framework is an analytical tool for judging whether the available data support a postulated mode of carcinogenic action.

Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies (OECD, 2002b). The Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies (GD 35), provides valuable information on assessment of the quality, integrity, and completeness of the experimental data from chronic toxicity and carcinogenicity studies and determining the acceptability of the study reports, including aspects such as reliability, relevance and adequacy (OECD, 2002b). GD 35 also provides guidance on the reporting of the study and on the use of historical control data as an adjunct to the internal controls of the study in question. These aspects are therefore considered to lie outside the scope of the current Guidance Document, and GD 35 should be consulted for further information.

1.2 Scope of application of the guidance

11. The Test Guidelines 451, 452 and 453 and this guidance document are designed to be used in the testing of a wide range of chemicals, whatever their field of application, including pesticides, industrial chemicals and pharmaceuticals. However, as noted in the Test Guideline 451, some testing requirements may differ for pharmaceuticals. The International Conference on Harmonisation of Testing for Pharmaceuticals (ICH) has produced a series of safety guidelines for the testing of pharmaceuticals, including guidelines on toxicokinetics (ICH, S3A), genotoxicity testing (ICH, S2), duration of chronic toxicity testing in animals (rodent and non-rodent toxicity testing) (ICH, M3(R2)), testing for carcinogenicity of pharmaceuticals (ICH, S1B) and dose selection for carcinogenicity studies of pharmaceuticals (ICH, S1C(R2)). These guidelines should always be consulted for specific guidance when testing pharmaceuticals using the approaches outlined in TG 451, 452 and 453. This guidance document provides, in various sections, examples of where the testing requirements may be different for pharmaceuticals.

1.3 Objectives of a chronic toxicity study

12. The objective of a chronic toxicity study, such as described by TG 452, is to characterize the toxicological response of a substance in a mammalian species following prolonged and repeated exposure. The chronic toxicity study provides information on the possible health hazards likely to arise from repeated exposure over a prolonged period of time. Key objectives of the study are to provide information useful for classification and labelling and an estimate of a point of departure (e.g., lower confidence limit of the benchmark dose (BMDL) or the no-observed-adverse-effect level (NOAEL)) for any adverse effects, which can be used for establishing safety criteria for human exposure. The study will provide information on the major toxic effects, and indicate target organs, progressive toxic effects and the possibility of delayed toxicity. The need for careful clinical observations of the animals, so as to obtain as much information as possible, is also stressed. Previous repeated dose 28-day and/or 90-day toxicity tests on a chemical may, among others, have indicated the potential to cause neurotoxic/neurobehavioural effects, or effects on the endocrine system, warranting further in-depth investigation as part of a chronic toxicity study.

1.4 Objectives of a carcinogenicity study

13. The objective of a long-term carcinogenicity study, such as described by TG 451, is to observe test animals for a major portion of their life span for the development of neoplastic lesions during or after exposure to various doses of a test substance by an appropriate route of administration. The carcinogenicity study may also provide information on the possible health hazards likely to arise from repeated exposure for a period lasting up to the entire lifespan of the species used. The study will provide information on potential carcinogenicity, and may indicate toxic effects, target organs, progressive toxic effects and the possibility of delayed toxicity. It can provide information useful for classification and labelling, and an estimate of a point of departure for toxic effects and, in the case of non-genotoxic carcinogens, for tumour responses, which can be used for establishing safety levels for human exposure. The need for careful clinical observations of the animals, so as to obtain as much

information as possible, is also stressed. Such an assay requires careful planning and documentation of the experimental design, a high standard of pathology, and unbiased statistical analysis. These requirements are well known and have not undergone any significant changes in recent years.

1.5 Objectives of a combined chronic toxicity/carcinogenicity study

14. The objective of a combined chronic toxicity/carcinogenicity study, such as described by TG 453, is to determine the toxicological effects of a substance (including carcinogenic potential) in a mammalian species following prolonged and repeated exposure. The combined chronic toxicity/carcinogenicity study provides information on the possible health hazards likely to arise from repeated exposure over the majority of the entire lifespan (in rodents). The study will provide information on the major toxic effects of the substance including potential carcinogenicity, and indicate target organs, progressive toxic effects and the possibility of delayed toxicity. The need for careful clinical observations of the animals, so as to obtain as much information as possible, is also stressed. The application of the TG 453 should generate data on which to identify the majority of chronic and carcinogenic effects and to determine dose-response relationships. Ideally, the design and conduct should allow for the detection of neoplastic effects and a determination of carcinogenic potential as well as general toxicity, including neurological, physiological, biochemical, and haematological effects and exposure-related morphological (pathology) effects.

15. The design of the updated TG 453 recommends, for the chronic phase of the study, at least three dose groups and a control group, each group containing at least 10 males and 10 females per group. The reduction of the number of animals per sex in the updated TG 453 compared to the initial version (1981) is justified on the basis of further information being available from animals on the carcinogenicity phase of the study and a more careful use of animals in laboratories than in 1981. The design of the carcinogenicity phase of the revised TG 453 is identical to the revised TG 451. The study will thus provide similar information on chronic toxicity and carcinogenicity as TG 452 and TG 451. It will allow derivation of a point of departure (e.g., BMDL or NOAEL), and will offer greater efficiency in terms of time and cost compared to conducting two separate studies, without compromising the quality of the data in either the chronic phase or the carcinogenicity phase. The data from both phases (chronic toxicity and carcinogenicity) will reinforce each other, as the animals used in the studies are drawn from the same stock and have similar characteristics at the start of the study. Measurements carried out on the animals in one phase will be relevant for the animals in the other phase, e.g., clinical signs, body weights, haematology and biochemistry (if carried out), pathology. The terminal kill of the chronic phase can act as an interim kill for the carcinogenicity phase.

1.6 Consideration of testing strategies

16. This section refers to testing strategies that may be applicable for certain regulatory authorities but have not been formally adopted by all. It does not recommend any particular testing strategy or approach, but suggests consideration of such approaches as part of an ongoing strategy to assess the toxic potential of a substance in an intelligent and iterative manner. As new validated approaches become scientifically appropriate for use in chronic toxicity or carcinogenicity assessment, and accepted by the relevant regulatory authorities, the study sponsor is encouraged to implement them where possible.

17. A reasoned scientific approach to the assessment of substances for chronic toxicity or carcinogenicity must first include an assessment of all available information that has the potential to influence the study design. This can include the identity, molecular structure, class, and physico-chemical properties of the test substance; any information regarding mode of action; results of relevant *in vitro* or *in vivo* toxicity tests such as genotoxicity, subchronic toxicity and toxicokinetics studies; anticipated use(s) and potential for human exposure; available (Q)SAR data; and relevant

toxicological data on structurally-related substances. This analysis can focus the study parameters, but may also lead to the conclusion that a study can be refined in some way, or not conducted at all based on a weight of evidence (Carmichael *et al.*, 2006; Doe *et al.*, 2006; Barton *et al.*, 2006; Cooper *et al.*, 2006).

18. Integrating a wide range of information to determine the potential toxicity of a substance is becoming more common as the gap between assessments that need to be conducted and the resources with which to conduct such assessments widens. Efforts are underway in many OECD countries to determine ways in which assessments of substances can be satisfactorily completed, and protection of public health and the environment achieved, while minimising costs in terms of time, money and animal use. However, the acceptability and use of testing strategies and weight-of-evidence approaches differ among OECD countries and regulatory sectors; thus, application of these approaches should always occur in consultation with appropriate regulatory authorities.

19. Shorter-term *in vitro* or *in vivo* tests may provide information regarding potency, mode of action, metabolism, and/or target organ that can help refine the chronic toxicity study protocol parameters or priorities for observation. Tiered approaches using a combination of *in silico*, *in vitro*, and *in vivo* tests have been proposed but are not yet widely implemented (Worth and Balls, 2002; Becker *et al.*, 2007).

20. A phased or tiered approach to the assessment of the carcinogenic potential of a substance should also be considered (Ashby, 1996). A number of shorter-term tests can be conducted which will provide useful information for determining whether and how a substance may be carcinogenic, including genetic toxicity assays, cell transformation or other cell-based assays, short-term cancer initiation-promotion tests which may or may not include toxicogenomic analyses (Ellinger-Ziegelbauer *et al.*, 2005; 2008), and *in vivo* repeated dose 28- or 90-day toxicity tests (for a review on *in vitro* and *in vivo* short term test see Maurici *et al.*, 2005). (Q)SAR prediction models have been used in a regulatory context to predict the carcinogenic potential of chemicals for several decades. There are also commercial (Q)SAR models available for predicting rodent carcinogenicity (Benigni *et al.*, 2007). (Q)SAR models should be validated according to OECD principles (OECD 2007, GD No. 69).

21. A number of different strategies for assessing carcinogenicity have been proposed (Langley 2001; Worth and Balls 2002; Knight *et al.*, 2006; Combes *et al.*, 2008). All feature a stepwise process or decision tree that prescribes information analysis and stopping points where classification and labelling and/or risk assessment could be possible. However, specific approaches have not yet been optimised or validated.

22. Consideration of particular tests or approaches should always be made within the context of whether the results will contribute mechanistic information that will be useful in the weight-of-evidence assessment of carcinogenic potential (OECD, 2002b).

23. The US National Toxicology Program, along with institutes in other OECD countries, has had a longstanding interest in the use of transgenic or knockout mouse models for the assessment of carcinogenicity (Bucher and Portier, 2004), as they consider that these models offer potential refinements, in terms of study duration and animal numbers, over the traditional long-term bioassay. At the time this guidance was prepared, some regulatory authorities in the pharmaceutical sector may accept studies with these models in combination with a full long-term rat bioassay in lieu of a second full bioassay in mice (ICH, 1997). In the past, the predictive ability of the models, and any refinements or animal reductions, has been questioned (Goodman, 2001; van Zeller and Combes, 1999; RIVM, 2004). A detailed review paper (DRP) on transgenic rodent mutation assays prepared by Canada was published by OECD (OECD, 2009).

1.7 Animal welfare considerations

24. The principles of the "3Rs" (Replacement, Reduction, and Refinement), first articulated by Russell and Burch in 1959 (Russell and Burch, 1959), should be considered as integral to the assessment of carcinogenicity or chronic toxicity in mammals, in order to ensure sound science, maximize animal welfare, and minimize animal use. Animals in a condition of stress or distress have a documented effect on the outcome of the study (Olsson and Dahlborn, 2001; Reinhardt and Reinhardt, 2002; Hurst and West, 2010). For these reasons the following principles should be implemented as much as practicably possible.

25. First and foremost, as discussed above, consideration of documented existing information from any reliable source that could provide a refinement in the testing protocol or procedure is recommended. Existing information could be used to inform dose spacing or selection, exposure route, observation priorities, potential modes of action or target organs of the test substance, and/or study design. Use of this information to focus the study before it begins ensures that the study will meet the expectations of the study sponsor and/or regulatory authorities, decreasing the likelihood of repeat studies.

26. The use of the combined chronic toxicity/carcinogenicity study (TG 453) is also recommended, which can in most cases accomplish the objectives of both studies, and offers savings in the numbers of animals used. This is due to the use of 10 animals per sex per dose group for the chronic toxicity phase of the study instead of 20 animals per sex per dose group when the carcinogenicity and the chronic toxicity studies are performed separately.

27. Any studies involving animals should abide by the principles of humane euthanasia as detailed in the OECD Guidance Document 19 on the recognition, assessment, and use of clinical signs as humane endpoints for experimental animals used in safety evaluation, and in particular paragraph 62 thereof (OECD, 2000). This paragraph states that "In studies involving repeated dosing, when an animal shows clinical signs that are progressive, leading to further deterioration in condition, an informed decision as to whether or not to humanely kill the animal should be made. The decision should include consideration as to the value of the information to be gained from the continued maintenance of that animal on study relative to its overall condition. If a decision is made to leave the animal on test, the frequency of observations should be increased, as needed. It may also be possible, without adversely affecting the purpose of the test, to temporarily stop dosing if it will relieve the pain or distress, or reduce the test dose." Close and frequent observations are recommended in order to determine the status of the animals, and any animals exhibiting clear signs of severe pain or distress should be humanely killed.

28. Animals may be housed individually, or be caged in small groups of the same sex; individual housing should be considered only if scientifically justified. Further detailed information on housing, feeding, and handling will be provided in Section 3.5.

29. As will be further discussed in Section 3.2, while the route of administration will depend on the physical and chemical characteristics of the test substance and expected route of human exposure, mixing the test substance into the diet or water is normally recommended for rodent studies. Administration of the test substance by oral gavage in carcinogenicity and chronic toxicity testing is normally not recommended for the reasons outlined in Section 3.2. If the oral gavage route is employed then its use should be justified. The testing of substances at potentially irritating or corrosive concentrations/doses should be avoided, as administering such substances could result in severe pain and tissue damage at point-of-entry, which would compromise both animal welfare and the integrity of the study.

30. Testing the chronic toxicity or carcinogenicity of inhaled substances can be achieved using either of two exposure conditions: whole-body or nose-only/snout-only. For studies of liquid or solid aerosols and for vapor that may condense to form aerosols, the nose-only exposure method allows the avoidance of oral exposure due to grooming of particles deposited on the fur. However, the welfare implications of a 1- or 2-year nose-only exposure study, and the potential for physiological effects of stress experienced by the animals to affect the results of the study, can lead to a preference for the use of the whole-body mode of exposure (Thomson *et al.,* 2009). Reasons for choice of exposure system should be justified in the study report. If rodent species other than rats are exposed nose-only, maximum exposure durations may be adjusted to minimise species-specific distress (GD 39).

31. Guidelines providing practical advice on animal welfare specifically for today's cancer researchers were updated in 2010. They define and encourage sharing of best practice in laboratory work in the field of cancer research (Workman *et al.*, 2010). Relevant information can be found in these guidelines, that also applies to toxicological studies.

REFERENCES

Ashby, J. (1996), "Alternatives to the 2-Species Bioassay for the Identification of Potential Human Carcinogens", *Hum Exp Toxicol,* Vol. 15: 183-202.

Barton, H.A., T.P. Pastoor, T. Baetcke, J.E. Chambers, J. Diliberto, N.G. Doerrer, J.H. Driver, C.E. Hastings, S. Iyengar, R. Krieger, B. Stahl and C. Timchalk (2006), "The Acquisition and Application of Absorption, Distribution, Metabolism, and Excretion (ADME) Data in Agricultural Chemical Safety Assessments", *Crit Rev Toxicol,* Vol. 36: 9-35.

Becker, R.A., L.M. Plunkett, J.F. Borzelleca and A.M. Kaplan (2007), "Tiered Toxicity Testing: Evaluation of Toxicity-Based Decision Triggers for Human Health Hazard Characterization", *Food and Chemical Toxicology,* Vol. 45: 2454–2469.

Benigni, R., T.I. Netzeva, E. Benfenati, C. Bossa, R. Franke, C. Helma, E. Hulzebos, C. Merchant, A. Richard, Y.T. Woo and C. Yang (2007), "The Expanding Role of Predictive Toxicology: an Update on the (Q)SAR Models for Mutagens and Carcinogens", *J Env Sci Hlth* Part C, Vol.25: 53-97.

Boobis, A.R., S.M. Cohen, V. Dellarco, D. McGregor, M.E. Meek, C. Vickers, D. Willcocks and W. Farland (2006), "IPCS Framework for Analyzing the Relevance of a Cancer Mode of Action for Humans", *Crit. Rev. in Toxicol.*, 36:793-801.

Bucher, J. and C. Portier (2004), "Human Carcinogenic Risk Evaluation, Part V: The NationalToxicology Program Vision for Assessing the Human Carcinogenic Hazard of Chemicals", *Tox Sci,* Vol. 82: 363–366.

Cohen, S.M., M.E. Meek, J.E. Klaunig, D.E. Patton and P.A. Fenner-Crisp (2003), "The Human Relevance of Information on Carcinogenic Modes of Action: An Overview", *Crit. Rev. Toxicol.,* Vol. 33:581-589.

Cooper, R.L., J.S. Lamb, S.M. Barlow, K. Bentley, A.M. Brady, N.G. Doerrer, D.L Eisenbrandt, P.A. Fenner-Crisp, R.N. Hines, L.F. Irvine, C.A. Kimmel, H. Koeter, A.A. Li, S.L. Makris, L.P. Sheets, G. Speijers and K.E. Whitby (2006), "A Tiered Approach to Life Stages Testing for Agricultural Chemical Safety Assessment", *Crit Rev Toxicol,* Vol. 36: 69-98.

Carmichael, N.G., H.A. Barton, A.R. Boobis, R.L. Cooper, V. Dellarco, P.A. Fenner-Crisp, J.E. Doe, J.C. Lamb and T.P. Pastoor (2006), "Agricultural Chemical Safety Assessment: A Multisector Approach to the Modernization of Human Safety Requirements", *Crit Rev Toxicol,* 36: 1-7.

Combes, R., C. Grindon, M.T.D. Cronin, D.W. Roberts and J.F. Garrod (2008), "Integrated Decision-tree Testing Strategies for Mutagenicity and Carcinogenicity with Respect to the Requirements of the EU REACH Legislation", *Alt Lab Anim,* Vol. 36: 43-63.

Doe, J.E., A.R. Boobis, A. Blacker V. Dellarco, N.G. Doerrer, C. Franklin, J.I. Goodman, J.M. Kronenberg, R. Lewis, E.E. Mcconnell, T. Mercier, A. Moretto, C. Nolan, S. Padilla, W. Phang, R. Solecki, L. Tilbury, B. van Ravenzwaay and D.C. Wolf (2006), "A Tiered Approach to Systemic Toxicity Testing for Agricultural Chemical Safety Assessment", *Crit Rev Toxicol,* Vol. 36: 37-68.

Ellinger-Ziegelbauer, H., H. Gmuende, A. Bandenburg and H.J. Ahr (2008), "Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term *in vivo* studies", *Mutat Res,* Vol. 637: 23-39.

Ellinger-Ziegelbauer, H., B. Stuart, B. Wahle, W. Bomann and H.J. Ahr (2005), "Comparison of the expression profiles induced by genotoxic and non-genotoxic carcinogens in rat liver", *Mutat Res / Fundamental and Molecular Mechanisms of Mutagenesis,* Vol. 575: 61-84.

EPA (2005), *Guidelines for Carcinogen Risk Assessment,* Risk Assessment Forum, U.S. Environmental Protection Agency, Washington DC.

Goodman, J.I. (2001), "A Perspective on Current and Future Uses of Alternative Models for Carcinogenicity Testing", *Toxicol Pathol,* Vol. 29 (Suppl): 173-176.

Holsapple, M.P., H.C. Pitot, S.N. Cohen, A.R. Boobis, J.E. Klaunig, T. Pastoor, V.L. Dellarco and Y.P. Dragan (2006), "Mode of Action in Relevance of Rodent Liver Tumors to Human Cancer Risk". *Toxicol. Sci*., Vol. 89:51-56.

Hurst, J.L. and R.S. West (2010), "Taming Anxiety in Laboratory Mice", *Nature Methods*, 7(10), 825-826.

ICH (International Conference on Harmonisation) (2009), *ICH Harmonised Tripartite Guideline - Guidance on Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals – M3(R2)*. Available at: http://www.ich.org/cache/html/501-272-1.html.

ICH (International Conference on Harmonisation) (1994), *ICH Harmonised Tripartite Guideline - Note for Guidance on Toxicokinetics: The Assessment of Systemic Exposure in Toxicity Studies – S3A*. Available at: http://www.ich.org/LOB/media/MEDIA495.pdf.

ICH (International Conference on Harmonisation) (1997), *ICH Harmonised Tripartite Guideline - Testing for Carcinogenicity of Pharmaceuticals - S1B.*
Available at: http://www.ich.org/LOB/media/MEDIA490.pdf.

ICH (International Conference on Harmonisation) (1998), *ICH Harmonised Tripartite Guideline - Duration of Chronic Toxicity Testing in Animals (Rodent and Non-Rodent Toxicity Testing) – S4.*
Available at: http://www.ich.org/LOB/media/MEDIA497.pdf.

ICH (International Conference on Harmonisation) (2008), *ICH Harmonised Tripartite Guideline - Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use – S2.* Available at: http://www.ich.org/LOB/media/MEDIA4474.pdf.

ICH (International Conference on Harmonisation) (2008), *ICH Harmonised Tripartite Guideline - Dose Selection for Carcinogenicity Studies of Pharmaceuticals – S1C.*
Available at: http://www.ich.org/LOB/media/MEDIA491.pdf.

ILSI (International Life Sciences Institute) (1997), *Principles for the Selection of Doses in Chronic Rodent Bioassays,* Foran JA (Ed.), ILSI Press, Washington, DC.

Knight, A., J. Bailey and J. Balcombe (2006), "Animal Carcinogenicity Studies: 3. Alternatives to the Bioassay", *Alt Lab Anim,* 34: 30-48.

Langley, G. (2001), *The way forward - Action to end animal toxicity testing*, London, UK: BUAV.

Maurici, D., M. Aardema, R. Corvi, M. Kleber, C. Krul, C. Laurent, N. Loprieno, M. Pasanen, S. Pfuhler, B. Phillips, D. Prentice, E. Sabbioni, T. Sanner and P. Vanparys (2005), "Carcinogenicity", *Alt Lab Anim,* Vol. 33 (Suppl 1): 177-182.

Meek, E.M., J.R. Bucher, S.M. Cohen, V. Dellarco, R.N. Hill, L.D. Lehman-McKemmon, D.G. Longfellow, T. Pastoor, J. Seed, and D.E. Patton (2003), "A Framework for Human Relevance analysis of Information on Carcinogenic Modes of Action", *Crit. Rev. Toxicol.,* Vol. 33:591-653.

OECD (2000), *Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation,* Series on Testing and Assessment N° 19, OECD, Paris.

OECD (2002a), *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies,* Series on Testing and Assessment N° 32, OECD, Paris.

OECD (2002b) *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies,* Series on Testing and Assessment N° 35, OECD, Paris.

OECD (2007), *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(QSAR)] Models*, Series on testing and assessment N° 69, OECD Paris.

OECD (2009), *Detailed Review Paper on Transgenic Rodent Mutation Assays*, Series on testing and assessment N° 103, OECD Paris.

OECD (2009), *Carcinogenicity Studies*, OECD Guideline for the Testing of Chemicals, No. 451, OECD, Paris.

OECD (2009), *Chronic Toxicity Studies,* OECD Guideline for the Testing of Chemicals, No. 452, OECD, Paris.

OECD (2009), *Combined Chronic Toxicity/Carcinogenicity Studies,* OECD Guideline for the Testing of Chemicals, No. 453, OECD, Paris.

Olsson, A. and K. Dahlborn (2001), "Improving housing conditions for laboratory mice: a review of environmental enrichment", *Lab Animals,* Vol. 36: 243-270.

Reinhardt, V. and A. Reinhardt (2002), *Comfortable Quarters for Laboratory Animals*, Animal Welfare Institute, Washington, D.C.

Rhomberg, L., K. Baetcke, J. Blancato, J. Bus, S. Cohen, R. Conolly, R. Dixit, J. Doe, K. Ekelman, P. Fenner-Crisp, P. Harvey, D. Hattis, A. Jacobs, D. Jacobson-Kram, T. Lewandowski, R. Liteplo, O. Pelkonen, J. Rice, D. Somers, A. Turturro, W. West and S. Olin (2007), "Issues in the Design and Interpretation of Chronic Toxicity and Carcinogenicity Studies in Rodents: Approaches to Dose Selection" *Crit. Rev. Toxicol.,*Vol. 37:729-837.

RIVM (2004), *Transgenic Mice as Alternative in Carcinogenicity Testing: Current Status*, RIVM report 340700001/2004, Netherlands.
Available at: http://rivm.openrepository.com/rivm/bitstream/10029/8911/1/340700001.pdf

Russell, W.M.S. and R.L. Burch (1959), *The Principles of Humane Experimental Technique*, Methuen Publishing Ltd., London.

Sonnich-Mullin, C., R. Fielder, J. Wiltse, K. Baetcke, J. Dempsey, P. Fenner-Crisp, D Grant, M. Hartley, A. Knaap, D. Kroese, I. Mangelsdorf, E. Meek, J. Rice, and M. Younes (2001), "IPCS Conceptual Framework for Evaluating a Mode of Action for Chemical Carcinogenesis" *Reg. Toxicol. & Pharm.,* Vol. 34:146-152.

Thomson, E.M., A. Williams, C.L. Yauk and R. Vincent (2009), "Impact of Nose-Only Exposure System on Pulmonary Gene Expression", *Inhal Toxicol,* 21(S1): 74-82.

Van Leeuwen, C.J., G.Y. Patlewicz and A.P. Worth (2007), *Intelligent Testing Strategies in Risk Assessment of Chemicals, an Introduction*, 2nd Edition, C.J. van Leeuwen and T.G.Vermeire, eds. Springer Netherlands, Rotterdam, pp 467-509.

Van Zeller, A.M. and R.D. Combes (1999), "Transgenic Mouse Bioassays for Carcinogenicity Testing: A Step in the Right Direction?", *Alt Lab Anim,* Vol. 27: 39–846.

Workman, P., E.O. Aboagye, F. Balkwill, A. Balmain, G. Bruder, D.J. Chaplin, J.A. Double, J. Everitt, D.A.H. Farningham, M.J. Glennie, L.F. Kelland, V. Robinson, I.J. Stratford, G.M. Tozer, S. Watson, S.R. Wedge and S.A. Eccles (2010), "Guidelines for the Welfare and Use of Animals in Cancer Research", *British Journal of Cancer,* 102, 1555–1577.

Worth, A. and M. Balls (2002), "Alternative (Non-animal) Methods for Chemicals Testing: Current Status and Future Prospects. A Report prepared by ECVAM and the ECVAM Working Group on Chemicals", *Alt Lab Anim,* Vol. 30 (Suppl 1): 1-125.

2. MODE OF TOXICOLOGIC ACTION

2.1 Introduction

32. This chapter provides an overview of the mode of action framework that has been developed to test hypothesized pathways of carcinogenicity and toxicity in recent years. It is based largely on the U.S. EPA Guidelines for Carcinogen Risk Assessment, (USEPA, 2005) and the International Programme on Chemical Safety (IPCS) Workshop Report (IPCS, 2005). It is provided primarily for the information of those carrying out long-term bioassays, rather than as specific guidance on determination of mode of action. It is important to note that while this chapter focuses primarily on the cancer endpoint, the mode of toxicologic action may be more generically applied to both cancer and noncancer endpoints.

33. The integrative approach to understanding of a chemical's toxicologic pathway, by weighing data from short-term toxicity, subchronic toxicity, genotoxicity and toxicokinetic data, becomes critical when designing a long-term bioassay. If a mode of action can be proposed, this information could greatly enhance the design of the long-term bioassay, in order to offer better insight into how a chemical elicits its toxicity and to help elucidate the shape of the dose-response curve. To fully utilize information on the hypothesized mode of action, the design of the study needs to account for doses that need to be placed carefully so as to yield observations of subtle precursor effects or other biomarkers of toxicity without inducing confounding effects related to frank toxicity (see also Section 3.1, Dose Selection).

2.2 Background

34. Animal long-term bioassays have been used for more than a half century to determine whether pesticides, pharmaceuticals, industrial chemicals, and other types of substances might cause cancer or other adverse chronic health problems in humans. As such, cancer bioassays, recognized by national and international regulatory groups, have become the default for testing the carcinogenic potential of products when human use or exposure is anticipated. Inherent in these animal-based assessments is the assumption that the observation of tumours in animals is directly relevant to the risk of cancer in humans, and that the responses observed at high doses in animals could be meaningfully extrapolated to doses with relevance for humans (IPCS, 2005). In order to predict responses in people more accurately, information is preferred from animals that are as similar to people as possible. More specifically, the use of other mammals, such as dogs, rats, and mice, as models for responses in humans is based on the assumption that there are important similarities among mammals in the way they respond to chemicals. In fact, a qualitative similarity has been established in the response of laboratory animals and humans to carcinogenic substances. Most known human carcinogens have been shown to be positive for tumourigenicity in well-conducted animal studies (Parekh and Dearfield, 2007).

35. Based on this assumption, such animal to human extrapolations, (dose and species) while necessary and practical, have been surrounded by intense discussion and debate. Through the evaluation of key events on the molecular and cellular level, a clearer understanding of how chemicals induce neoplasia has been realized. Such mechanistic data have called into question the appropriateness of extrapolating certain rodent tumour responses to humans. Thus, cancer risk assessments are rarely without controversy and are often heatedly debated within the scientific community (Holsapple *et al.*, 2005).

36. More recently, the understanding of the pathogenesis of neoplasia has evolved significantly. It is now recognized that cancers originating from at least some cell types may arise by a variety of independent pathways. It is also established that different carcinogens may have different modes of

action and that some carcinogens act through more than one mode of action in different tissues. While some modes of action lead to cancers in both rodents and humans, others that are carcinogenic in rodents are not in humans, at least under realistic circumstances of human exposure. To refine and improve the process of carcinogenic hazard identification, and to avoid misidentification of non-tumourigenic compounds as possible human carcinogens, it has become crucial that mode of action analysis be undertaken and that data to support such analysis be collected in a thorough and scientifically rigorous manner (Rice, 2004).

37. Risk assessments have benefited from our understanding of the mode of action of carcinogenesis in both animals and in humans and from the use of pharmacokinetic and pharmacodynamic data to determine the appropriateness of assumptions and to characterize the biological basis underlying the use of such assumptions. There has been increased recognition of advancements in scientific thinking on cancer that is reflected in some regulatory agency's adoption of the mode of action paradigm (USEPA, 2005). In accordance with this line of thinking, some regulatory agencies have also provided a systematic framework to test hypothesized toxicity pathways. Because of the many benefits of using this mode of action framework, implementation has been widespread and is now commonly used by additional regulatory agencies and international organizations (Meek *et al.*, 2003; Boobis *et al.*, 2006). In the United Kingdom, the mode of action framework is being applied to the assessments of pesticides and industrial chemicals. The UK Committee on Carcinogenicity (COC) has noted its value with regard to both harmonization between agencies and internal consistency in its latest Guidelines (COC, 2004). It has also been adopted and is being implemented by Canadian agencies, for example in the evaluation of Existing Chemicals under the Canadian Environmental Protection Act. The European Union has incorporated the framework into the technical guidance of assessments that are being updated for toxic effects, including carcinogenicity, of new and existing industrial chemicals (ECHA, 2009). The framework has been featured by the WHO/FAO Joint Meeting on Pesticide Residues in its evaluation of pyrethrin extract and its incorporation into the resulting monograph. Finally, IPCS in cooperation with international partners has taken steps to move the framework forward by melding it with the human relevance concept (IPCS Workshop, 2005).

2.3 "Mode of Action" Framework

38. The term "mode of action" is defined as a sequence of key events and processes, starting with interaction of an agent with a cell, proceeding through operational and anatomical changes, and resulting in e.g., cancer formation. A "key event" is an empirically observable precursor step that is a necessary element of the mode of action or is a biologically based marker for such an element. The term "mechanism of action" implies a more detailed understanding and description of events, often at the molecular level, than is meant by mode of action. The toxicokinetic processes that lead to formation or distribution of the active agent to the target tissue are considered in estimating dose but are not part of the mode of action as the term is used here. This chapter focuses on the carcinogenic mode of action. There are many examples of possible modes of carcinogenic action, such as mutagenicity, mitogenesis, inhibition of cell death, cytotoxicity with reparative cell proliferation, and immune suppression.

39. Elucidation of a mode of action for a particular carcinogenic response in animals or humans is a data rich determination. Significant information should be developed to ensure that a scientifically justifiable mode of action underlies the process leading to cancer at a given target site. In the absence of sufficiently scientifically justifiable mode of action information, regulatory scientists generally take a public health protective default position regarding interpretation of toxicologic and epidemiologic data; animal tumour findings are judged to be relevant to humans and cancer risks are assumed to conform with low dose linearity.

Mode of Action Framework: Animal Tumours

40. The framework is intended to be an analytic tool for systematically judging whether available data support a mode of carcinogenic action hypothesized for an agent in a transparent manner. It is not designed to give an absolute answer on sufficiency of the information as this will vary depending on the circumstance (IPCS, 2005). Amongst the strengths of the framework are its flexibility, general applicability to carcinogens acting by any mechanism and the ability to explore the impact of each key event on the carcinogenic response (IPCS, 2005). It is primarily based upon considerations for causality in epidemiologic investigations originally articulated by Hill (1965) but later modified by others and extended to experimental studies. The modified Hill criteria are useful in organizing thinking about aspects of causation, and they are consistent with scientific method of developing hypotheses and testing those hypotheses experimentally. A key question is whether the data to support a mode of action meet the standards generally applied in experimental biology regarding inference of causation.

Components of a Mode of Action Analysis

41. To perform a mode of action analysis, the key biochemical, cellular and molecular events need to be established, and the temporal and dose-dependent concordance of each of the key events in the mode of action can then be determined. The key events can be used to bridge species and dose for a given mode of action. The next step in the mode of action analysis is the assessment of biological plausibility for determining the relevance of the specified mode of action in an animal model for human cancer risk based on kinetic and dynamic parameters. (Holsapple *et al.*, 2005)

Postulated Mode of Action: Key Events

42. The postulated mode of action is a biologically plausible hypothesis/basis for the sequence of key events leading to an observed effect supported by robust experimental observations and mechanistic data (IPCS, 2005). Key events are critical to the induction of tumours as hypothesized in the postulated mode of action. To support an association, a body of experiments needs to define and measure an event consistently.

43. To evaluate whether a hypothesized or postulated mode of action is operative, an analysis starts with an outline of the scientific findings regarding the hypothesized key events leading to cancer, and then weighing information to determine whether there is a causal relationship between these events and cancer formation. Again, it is not generally expected that the complete sequence will be known at the molecular level. Instead, observations made at different levels of biological organization (e.g., biochemical, cellular, physiological, tissue, organ, and system) are analyzed.

44. For each tumour site being evaluated, the mode of action analysis should begin with a description of the relevant data and key events that may be associated with a hypothesized mode of action and its sequence of key events. This can be followed by discussion of various aspects of the experimental support for hypothesized modes of action in animals and humans. (See Appendix I. for examples of mode(s) of action.)

Experimental Support for the Postulated Mode of Action

45. Experimental support addressing the strength, consistency and specificity of association, dose response concordance, temporal relationship and if the mode of action is biologically plausible all add to establishing a clear mode of action.

1. Strength, Consistency and Specificity of Association

46. A statistically significant association between key events and tumour response, observed in well conducted studies is generally supportive of causation. Consistent observations in a number of such studies with differing experimental designs increase that support, because different designs may reduce unknown biases. Studies showing absence/reduction of carcinogenicity when the rate limiting event is reversed, blocked or diminished, are particularly useful tests of association. Conversely, if enhancement of rate limiting key events increases the tumour response, this evidence would also provide strong support for the postulated mode of action. Pertinent observations include tumour response and key events in the same cell type, sites of action biologically related to key event(s), and results from multistage studies and from stop/recovery studies (Boobis *et al.*, 2006). Specificity of the association without evidence of other modes of action also strengthens a causal conclusion. And while these factors provide additional confidence that the primary mode of action has been identified, conversely, a lack of strength, consistency and specificity of an association tends to weaken the overall causal conclusions for a particular mode of action.

2. Dose Response Concordance

47. If a key event and tumour endpoints increase with dose such that the key events forecast the appearance of tumours at a later time or higher dose, the shape of the dose/response curve could be revealed and a causal association can be strengthened. Dose-response associations of the key event with other precursor events can add further strength.

3. Temporal Concordance

48. If a key event is shown to be causally linked to tumourigenesis, it should precede tumour appearance. An event may also be observed contemporaneously or after tumour appearance; these observations may add to the strength of association but not to the temporal association. Pertinent observations include studies of varying durations observing the temporal sequence of events and development of tumours (see paragraphs 67 and 68 on precursor/key events for application to long-term studies).

4. Biological Plausibility and Coherence

49. The biological plausibility of any postulated mode of action in humans depends on a consideration of dose-effect and dose-response relationships (IPCS, 2005). The postulated mode of action and key events should be based on contemporaneous understanding of the biology of cancer. If the body of information under scrutiny is consistent with other chemical agents for which the hypothesized mode of action is accepted, the case is strengthened. Note: Because some modes of action can be anticipated to evoke effects other than cancer, the available toxicity database on noncancer effects can contribute to this evaluation.

Alternative Mode(s) of Action

50. The possibility of other modes of action should also be considered and discussed. If there is evidence for more than one mode of action, each mode should receive a separate analysis. Furthermore, different modes of action can operate in different dose ranges; for example, an agent can act predominately at lower doses where cytotoxicity may not occur. Ultimately, however, information on all modes of action should be integrated to better understand how and when each mode acts, and which modes may be of primary interest for exposure levels relevant to human exposure of interest.

Uncertainties, Inconsistencies and Data Gaps

51. Uncertainties should be stated clearly, fully and explicitly. They should include those related to the biology of the toxicological response and those for the database on the specific chemical being evaluated. Any inconsistencies should be clearly noted and characterized with respect to the impact on the weight of evidence in support of the postulated mode action. Data gaps should also be identified and characterized. It should be clearly stated whether the identified data gaps are critical in supporting the postulated mode of action and what recommendations can be provided to address those data deficiencies in the future (Boobis *et al.*, 2008).

Conclusion of Postulated Mode of Action Analysis

52. Conclusions about each postulated mode of action should address (1) whether the mode of action is supported in animals, (2) whether it is relevant to humans and (3) which populations or life stages can be particularly susceptible. Special attention should be paid to whether tumours can arise from childhood exposure, considering various aspects of development during these life stages. Because the cancer studies are usually performed with young adult or juvenile animals, conclusions about relevance during early childhood generally rely on inference.

Relevance of rodent mode of action for humans

53. "Relevance" of a potential mode of action is considered in the context of characterization of hazard and not at the level of risk. Anticipated levels of human exposure are not used to determine whether the postulated mode of action is operative in a particular population or life stage, for example, in those with pre-existing disease. (USEPA, 2005) Human relevance is discussed in the following section (section 2.4), in the context of the Human Relevance Framework.

54. Other populations or life stages may not be analogous to the test animals, in which case the question of relevance would be decided by inference. Although agent specific data would be preferable, this review may also rely on general knowledge about the precursor events and characteristics of individuals susceptible to these key precursor events. Any information suggesting quantitative differences between populations or life stages should be flagged for consideration in the dose-response assessment, and a separate risk estimate should be quantified for susceptible populations or life stage if data suggests a quantitative difference.

2.4 Human Relevance Framework

55. Considerable effort has been expended during the past several decades to evaluate the mode of action for specific chemicals causing cancer in rodents. However, the key question is the relevance of this postulated mode of action to human risk assessment. A framework was developed by an ILSI/RSI working group sponsored by the U.S. EPA and Health Canada to address this issue and to provide direction in determining the relevance of rodent tumours to human health (Cohen *et al.*, 2003; Meek *et al.*, 2003; Cohen *et al.*, 2004). This human relevance framework is not prescriptive and does not provide a check list of criteria; it is an analytical tool that describes methods and a decision tree logic to establish a relationship between early cellular events and the development of cancer and its relevance to humans. Knowledge of key events and the identification of a mode of action provide a more rational basis for human hazard and risk assessment.

56. The human relevance framework is based on three questions: (1) is the weight of evidence sufficient to establish the mode of action in animals? (2) are key events in the animal mode of action plausible in humans? And (3) taking into account kinetic and dynamic factors, are key events in the

animal mode of action plausible in humans? This is a more quantitative analysis which addresses the relevance of tumourigenicity to a level of exposure, and again relies on a concordance analysis between animal model and humans. This approach focuses not only on dose response but also on quantitative differences between species in fundamental biologic processes that can affect exposure.

57. Presentation in tabular form referred to as a concordance table can be particularly useful. The information in these tables should be relatively brief, as a narrative explanation. There should be one column for the effect on humans for each key events evaluated and another column for the results in a different strain, species, or sex or for a different route of administration that does not result in toxicity. Factors may be identified that are not key events but can modulate key events and contribute to differences between species or individuals. Examples include genetic differences in pathways of metabolism, competing pathways of metabolism, and effects induced by concurrent pathology. While information for evaluating key events in humans may come from *in vitro* and *in vivo* studies on the chemical, basic information on anatomy, physiology, endocrinology, genetic disorders, human epidemiology, and other information that is known regarding the key events should be considered in this framework (Boobis *et al.*, 2008) .

58. This human relevance framework is focused on hazard identification and evaluation. If the second and third questions are answered in the negative, then there is not a cancer hazard for humans and therefore no cancer risk (Holsapple *et al.*, 2005). It is clearly acknowledged that departure from the default assumption of human relevance is a data rich determination. If a conclusion is strongly supported by empirical data, exposure to chemicals producing the toxicity only via that mode of action would not pose a risk to humans. Therefore, no additional risk characterization for this endpoint of carcinogenicity is further warranted (Boobis *et al.*, 2008).

59. Appendix II includes an example where Pastoor *et al.* (2005) describe a rodent mode of action they judge not relevant for humans. This determination for thiamethoxam-related mouse liver tumours was based on the quantitation of key metabolites *in vivo* and *in vitro* that showed mice, but not rats or humans to be capable of generating sufficient amounts of these metabolites to initiate the hepatic toxicity necessary for tumour formation.

Hazard Characterization

60. The hazard characterization provides the overall weight of evidence summary of the assessment. It summarizes the conclusions about the agent's potential effects, whether they can be expected to depend qualitatively on the circumstances of exposure, and if anyone can be expected to be especially susceptible. It discusses the extent to which these conclusions are supported by data or are the result of default options invoked because the data are inconclusive. It explains how complex cases with differing results in different studies were resolved. The hazard characterization highlights the major issues addressed in the hazard assessment and discusses alternative interpretations of the data and the degree to which they are supported scientifically.

61. When the conclusion is supported by mode of action information, the hazard characterization also provides a clear summary of the mode of action conclusions, including the completeness of the data, the strengths and limitations of the inferences made, the potential for other modes of action, and the implications of the mode of action for selecting viable approaches to the dose response assessment. The hazard characterization also discusses the extent to which mode of action information is available to address the potential for disproportionate risks in specific populations or lifestages or the potential for enhanced risks on the basis of interactions with other agents or stressors.

2.5 Consideration of "Mode of Action" information to Optimise the Design of Long-term Studies

An "Integrative" Approach

62. Before embarking on the design of a long-term chronic rodent study, the objective or goal of the study must be clearly defined. If the objective is to determine a chemical's carcinogenic potential, and understand how it elicits its carcinogenicity (mode of toxicological action), a weight of evidence, integrative approach needs to be considered. It is critical to explore potential challenges that might arise from having multiple objectives for the bioassay. For practical reasons, it is often the case that there are multiple objectives in this one study design. To this end, multiple objectives can be accomplished by considering how the design of the study could be optimized for each individual objective. This may necessitate the use of additional animals in the study, compared with the standard design of 50 animals of each sex per group laid down in the Test Guideline TG 451. However such use of additional animals in the study design may be justified if it can avoid the need for a separate study, involving the use of more animals overall. The use of additional animals or the introduction of new experimental groups must however be scientifically justified.

63. Using an integrative approach, the following information on the chemical would be weighed:

- What are the basic physicochemical properties of the chemical? Can the compound be administered orally for a long-term duration?
- Is the material a direct acting DNA mutagen; is it genotoxic?
- Does the chemical induce liver enzymes, liver weight increase, hypertrophy?
- How does the chemical causes its toxicity in shorter term studies and is this consistent across multiple species?
- Does the chemical cause hyperplasia or toxicity in particular target organs?
- Does the chemical cause cell proliferation, inflammation, cellular necrosis, apoptosis?
- Does the chemical cause hormonal perturbation in shorter term or subchronic animal studies? or
- Are there similar analogs, QSAR analyses or structural alerts for organ toxicity? Cancer?

All these data become relevant to the consideration of a chemical's primary mode of action and the determination of carcinogenic potential (Jacobs, 2005).

Hypothesized Mode of Action

64. If the objective of the study is to provide data to test hypotheses regarding a mode of toxicologic action based on information from a structurally similar analog, SAR prediction, or if a chemical "fits" structurally in a particular class of compounds that is known to have a specific mode of action and similar target organ, the study design should include the consideration of dose spacing, temporal sequence of key events specific to the proposed or suggested mode of action, and subsequent precursor key events. Some key questions include: what properties of the dose-response curve are most important, the steepness of the slope? How would one determine the placement of animals into the various dose categories and what are the necessary precursor events in the mode of action for carcinogenicity? The issues of a chemical's structural activity, dose selection and inclusion of precursor key events will be discussed in the ensuing paragraphs.

Consideration of Structure Activity Relationship (SAR)

65. If information on potential modes of action on the chemical of interest or similarly structured compound analog (e.g., SAR) is available, this could assist in identifying the possible, proposed mode of action. In general, the results of shorter term, subchronic studies can also provide information needed to identify necessary precursor events, and inform the selection of adequate doses for a subsequent carcinogenicity study. However, the dose levels administered in a subchronic study in the parent or structurally similar compound that induce a particular targeted effect may need to be adjusted (increased or decreased) for incorporation in a carcinogenicity study.

Dose Selection/Placement

66. Possible insights as to the postulated mode of action can be provided by obtaining information on how the toxicity of the chemical changes with increasing dose, which in turn can be provided by appropriate dose selection and placement (see also Section 3.1). This approach requires some previously generated information (e.g., shorter term studies or toxicokinetic data). If prior evidence allows, it may be possible to optimize the study design in terms of the location of the doses and the allocation of animals to the doses. Depending on the state of the scientific knowledge regarding possible mode(s) of action, sensitivity may be a less important constraint on dose group number than the need for an experimental design, that can compare the various alternative modes of actions. This is a method driven by hypotheses regarding the mode of action.

Precursor/Key Events

67. For mode of action determinations, there may be value in conducting additional, specific studies as a preliminary step to conducting a chronic/carcinogenicity study. With a hypothesized mode of action, the study design of the long-term bioassay can be modified to include additional precursor key events (enzyme induction, cellular proliferation, hormonal perturbation, necrosis and/or apoptosis, hyperplasia, foci development, clonal expansion, etc.) that would support a mode of action. This will normally involve the use of additional satellite animals in the study design, which can be justified if it can avoid the need for a separate study, involving the use of more animals overall. It may however be more appropriate to examine some of these key events in separate, shorter term, studies. Additional analyses of cellular proliferation at various time points (1, 7, 14, 28 days), or cellular necrosis and/or apoptosis and additional clinical chemistry parameters (e.g., gamma glutamyl transpeptidase (GGT), alanine transaminase (ALT), etc.) may be investigated, again in satellite groups or in a separate study, to provide a more robust mode of action analysis. In including these additional parameters in the study design of the long-term bioassay, it is important to consider dose selection and key events at doses lower than the top dose and observations occurring earlier than the time of appearance of the first tumours in the study.

68. Results from investigations applying non-standard test methods such as DNA-microarrays and other ‘omics’ tools, e.g., methods to globally analyze the expression of genes, proteins or metabolites may provide supplementary information to support the identification of primary targets and mechanisms on the molecular level. Characterization of key events and dose-related responses in a presumed toxic/carcinogenic mode of action may be used to predict the relevance of toxic effects observed in conventional tests used for human health-related end points. In a case-by case approach such data could contribute to the optimisation of the design of long-term studies. This way, the weight of evidence on the potential of the test substance to induce cancer can be evaluated. In order to ensure the validity of such non-standard test methods, any microscopic or measurable substance-related adverse effects of the conventional methods should correlate with the observed changes in expression of targeted genes, proteins or metabolites, taking into account the high sensitivity and appropriateness of the technique applied.

69. The use of information on mode of action and human relevance frameworks for interpretation of carcinogenicity studies in rodents is relatively new. Therefore Appendix I and Appendix II to this Chapter provide examples in which these approaches have been useful, in order to indicate the type of additional information that is valuable. These approaches should be taken into consideration in the design of a chronic toxicity or carcinogenicity study to ensure that the maximum information is obtained from the minimum number of animals.

REFERENCES

Boobis, A.R., S.M. Cohen, V. Dellarco, M.E. Meek, C. Vickers, D. Willcocks and W. Farland (2006), "IPCS Framework for Analyzing the Relevance of a Cancer Mode of Action for Humans", *Critical Reviews in Toxicology*, 36:781-792.

Boobis, A.R., J.E. Doe, B. Heinrich-Hirsch, M.E. Meek, S. Munn, M. Ruchirawat, J. Schlatter, J. Seed and C. Vickers (2008), "IPCS Framework for Analyzing the Relevance of a Noncancer Mode of Action for Humans", *Critical Reviews in Toxicology*, 38:87-96.

Cattley, R., J. DeLuca, C. Elcombe, P. Fenner-Crisp, B. Lake, D. Marsman, T. Oastoor, J.A. Popp, D.E. Robinson, B. Schwatz, J. Tugwood and W. Wahli (1998), "Do Peroxisome Proliferating Compounds Pose a Hepatocarcinogenic Hazard to Humans?" *Regul. Toxicol. Pharmacol*., 27:47-60.

Cohen, S.M., M.E. Meek, J.E. Klaunig, D.E. Patton and P.A. Fenner-Crisp, (2003), "The Human Relevance of Information on Carcinogenic Modes of Action: Overview", *Crit. Rev. Toxicol.* **33,** 581–589.

Cohen, S.M., J. Klaunig, M.E. Meek, R.N. Hill, T. Pastoor, L. Lehman-McKeeman, J. Bucher, D.G. Longfellow, J. Seed, V. Dellarco, P. Fenner-Crisp and D. Patton (2004), "Evaluating the Human Relevance of Chemically Induced Animal Tumors", *Toxicol. Sci.*, 78: 181-186.

Committee on carcinogenicity of chemicals (COC) in food, consumer products and the environment (2004), Guidance on a strategy for the risk assessment of chemical carcinogens. http://www.iacoc.org.uk/publications/documents/guideline04.pdf

Guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. http://guidance.echa.europa.eu/docs/guidance_document/clp_en.pdf

Ellinger-Ziegelbauer, H., B. Stuart, B. Wahle, W. Bormann and H.J. Ahr (2004), "Characteristic Eexpression Profiles Induced by Genotoxic Carcinogens in Rat Liver", *Toxicol. Sci.*, 77: 19-34.

Ellinger-Ziegelbauer, H., B. Stuart, B. Wahle, W. Bormann and H.J. Ahr (2005), "Comparison of the Expression Profiles Induced by Genotoxic and Nongenotoxic Carcinogens in Rat Liver", *Mutat. Res.* 575:61-84.

Hill, A.B. (1965), "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, 58:295–300.

Holsapple, M.P., H.C. Pitot, S.M. Cohen, A.R. Boobis, J.E. Klaunig, T. Pastoor, V.L. Dellarco and Y.P. Dragan (2006), "Mode of Action in Relevance of Rodent Liver Tumors to Human Cancer Risk", *Toxicol. Sci.*, 89:51-56.

IPCS (2005), *Record of the Cancer Framework Workshop, Bradford, United Kingdom, 21–23 April 2005.* Geneva, World Health Organization, International Programme on Chemical Safety.

Jacobs, A. (2005), "Prediction of 2-year Carcinogenicity Study Results for Pharmaceutical Products: How Are We Doing?" *Toxicological sciences*, 88(1), 18-23.

Meek, E.M., J.R. Bucher, S.M. Cohen, V. Dellarco, R.N. Hill, L.D. Lehman-McKemmon, D.G. Longfellow, T. Pastoor, J. Seed and D.E. Patton (2003), "A Framework for Human Relevance Analysis of Information on Carcinogenic Modes of Action", *Crit. Rev. Toxicol.,* 33:591-653.

Parekh, A.A. and K.L. Dearfield (2007), "USEPA's Risk Assessment Practice: Extrapolation Between Species", *Human and Ecological Risk Assessment*, 13:96-100.

Pastoor T., P. Rose, S. Lloyd, R. Peffer and T. Green (2005), "Case Study: Weight of Evidence Evaluation of the Human Health Relevance of Thiamethoxam-Related Mouse Liver Tumors", *Toxicological Sciences*, 86(1), 56-60.

Rhomberg L.R., K. Baetke, J. Blancato, J. Bus, S. Cohen, R. Conolly, R. Dixit, J. Doe, K. Ekelman, P. Fenner Crisp, P. Harvey, D. Hattis, A. Jacobs and D. Jacobson-Kram (2007), "Issues in the design and Interpretation of Chronic Toxicity and Carcinogenicity Studies in Rodents: Approaches to Dose Selection", *Critical Reviews in Toxicology*, 37:729-837.

Rice, J.M. (2004), "On the Application of Data on Mode of Action to Carcinogenic Risk Assessment", *Toxicol. Sci.*, 78, 175–177.

USEPA (2005), *Guidelines for Carcinogen Risk Assessmen,* Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC, USA.

EAP/630/P-03/001F http://cfpub.epa.gov/ncea/raf/recorddisplay.cfm?deid=116283

Ward, W., D. Delker, S. Hester, S.-F. Thai, D. Wolf, J. Allen and S. Nesnow (2006), "Transcriptional Profiles in Liver from Mice Treated with Hepatotumorigenic and Nonhepatotumorigenic Triazole Conazole Fungicides: Propiconazole, Triadimefon, and Myclobutanil", *Toxicol. Path.*, 34:863-878.

Appendix I. Examples of Animal Mode(s) of Action (MOA) Framework

1. Example(s) of Liver Cytotoxic Mode of Action: Chloroform and Carbon Tetrachloride

This paper summarizes recent developments in the continuing evolution of Human Relevance Frameworks to systematically consider the weight of evidence of hypothesized modes of action in animals and their potential human relevance for both cancer and non-cancer effects. These frameworks have been developed in initiatives of the International Life Sciences Institute Risk Sciences Institute and the International Programme on Chemical Safety engaging large numbers of scientists internationally. They are analytical tools designed to organize information in hazard characterization as a basis to clarify the extent of the weight of evidence for mode of action in animals and human relevance and subsequent implications for dose-response. They are also extremely helpful in identifying critical data gaps. These frameworks which are illustrated by an increasing number of case studies, have been widely adopted into international and national guidance and assessments and continue to evolve, as experience increases in their application (Meek, 2008).

Under the 2005 U.S. EPA Guidelines for Carcinogen Risk Assessment, evaluations of carcinogens rely on mode of action data to better inform dose response assessments. A reassessment of carbon tetrachloride, a model hepatotoxicant and carcinogen, provides an opportunity to incorporate into the assessment biologically relevant mode of action data on its carcinogenesis. Mechanistic studies provide evidence that metabolism of carbon tetrachloride via CYP2E1 to highly reactive free radical metabolites plays a critical role in the postulated mode of action. The primary metabolites, trichloromethyl and trichloromethyl peroxy free radicals, are highly reactive and are capable of covalently binding locally to cellular macromolecules, with preference for fatty acids from membrane phospholipids. The free radicals initiate lipid peroxidation by attacking polyunsaturated fatty acids in membranes, setting off a free radical chain reaction sequence. Lipid peroxidation is known to cause membrane disruption, resulting in the loss of membrane integrity and leakage of microsomal enzymes. By-products of lipid peroxidation include reactive aldehydes that can form protein and DNA adducts and may contribute to hepatotoxicity and carcinogenicity, respectively. Natural antioxidants, including glutathione, are capable of quenching the lipid peroxidation reaction. When glutathione and other antioxidants are depleted, however, opportunities for lipid peroxidation are enhanced. Weakened cellular membranes allow sufficient leakage of calcium into the cytosol to disrupt intracellular calcium homeostasis. High calcium levels in the cytosol activate calcium-dependent proteases and phospholipases that further increase the breakdown of the membranes. Similarly, the increase in intracellular calcium can activate endonucleases that can cause chromosomal damage and also contribute to cell death. Sustained cell regeneration and proliferation following cell death may increase the likelihood of unrepaired spontaneous, lipid peroxidation- or endonuclease-derived mutations that can lead to cancer. Based on this body of scientific evidence, doses that do not cause sustained cytotoxicity and regenerative cell proliferation would subsequently be protective of liver tumours if this is the primary mode of action. To fulfill the mode of action framework, additional research may be necessary to determine alternative mode(s) of action for liver tumours formed via carbon tetrachloride exposure (Manibusan *et al.*, 2007).

References

Manibusan, MK, M. Odin and D.A. Eastmond (2007), "Postulated Carbon Tetrachloride Mode of Action: a Review", *J Environ Sci Health C. Environ Carcinog Ecotoxicol Rev* Jul-Sep; 25(3), 185-209.

Meek, B. (2008), "Recent Developments in Frameworks to Consider Human Relevance of Hypothesized Modes of Action for Tumors in Animals", *Environmental and Molecular Mutagenesis*, Volume 49, Issue 2, 110-116.

SAB (1999), *Minutes of the Chloroform Risk Assessment Review Subcommittee Meeting*, October 27-28, 1999, U.S. Environmental Protection Agency, Science Advisory Board. See website: http://www.epa.gov/sab/chloro00.htm

2. Example of Urothelial Cytotoxicity and Increased Cellular Proliferation: Dimethylarsinic Acid

The dose-response relationship for Dimethyl Arsenic (DMA) tumourigenesis based on mode of action considerations will be nonlinear as it is dependent on genetic, biochemical and histopathological events for which dose-response relationships are nonlinear. There must be a sufficient concentration of DMAIII in the bladder to produce cell death and regenerative proliferation. The dose-response assessment would ideally be based on use of DMAIII dosimetry at the target tissue because it represents the rate-limiting event of reductive metabolism to DMAIII to provide a level of exposure that will be protective against the key event of regenerative proliferation. Therefore, the mode of action analysis shows that sufficient DMAIII must be present to result in sufficient urothelial cytotoxicity and cell killing to result in increase cell poliferation and associated chromosomal aberrations. All of these events must occur to result in a neoplastic response. Any one event alone is not sufficient to lead to tumours.

References

USEPA (2005), Science Issue Paper: Mode of Carcinogenic Action for Cacodylic Acid (Dimethylarsinic Acid, DMAV) and Recommendations for Dose Response Extrapolation July 26, 2005, Health Effects Division, Office of Pesticide Programs, US Environmental Protection Program. See website;
http://www.epa.gov/oppsrrd1/reregistration/cacodylic_acid/dma_moa.pdf

SAB (2007), Advisory on EPA's Assessments of Carcinogenic Effects of Organic and Inorganic Arsenic: A Report of the US EPA Science Advisory Board http://www.epa.gov/sab/pdf/sab-07-008.pdf

3. Example of a Neuroendocrine Mode of Action: Atrazine

In 2000, EPA presented a proposed MOA for atrazine to the FIFRA SAP which supported the Agency's approach. EPA described this MOA and the relevant cancer and reproductive toxicity data in the "Atrazine: Hazard and Dose-Response Assessment and Characterization" (FIFRA SAP, 2000a). In brief, upon high levels of exposure to atrazine, the release of gonadotropin releasing hormone (GnRH) from the hypothalamus is reduced, thereby lessening the afternoon pituitary luteinizing hormone surge in female Sprague Dawley rats. As a result, the estrus cycle lengthens. This, in turn, leads to increased estrogen levels and an increased incidence of mammary tumours in female Sprague Dawley rats. (SAP, 2000)

References

SAP (2000), SAP Report No. 2000-05, FIFRA Scientific Advisory Panel Meeting, June 27-29, 2000, Held at the Sheraton Crystal City Hotel, Arlington, Virginia, Atrazine Hazard and Dose Response Assessment and Characterization. See website:
http://www.epa.gov/oscpmont/sap/meetings/2000/june27/finalatrazine.pdf

4. Example of a Mutagenic Mode of Action: Chromium (VI)

Based on the findings from the National Toxicology Program (NTP) 2-year drinking water studies in rodents that Cr (VI) significantly increased the incidence of carcinomas in the small intestine of male and female mice, a weight-of-the-evidence (WOE) approach was applied to judge the mutagenic data of Cr (VI) relative to the induction of gene mutations and/or chromosome aberrations. The next step was to determine whether the induction of gene mutations and chromosome aberrations, which was seen across *in vitro* studies, in animals (mice and rats) and in humans, was an early key event in the carcinogenic process. The synthesis and critical analysis of these data were instrumental in establishing whether the genetic damage occurs early and at doses that are within the tumourigenic range. These two criteria are the hallmarks of the *Cancer Guidelines* MOA framework and were clearly supported by the Cr (VI) data. While this is largely a data-rich undertaking, most of the studies evaluated in this document were primarily designed for hazard identification. Consequently, studies that specifically address various aspects of the MOA framework analysis are not always available for review. Despite this limitation, the WOE approach taken with Cr (VI) demonstrates the utility of this strategy in identifying missing data and establishes the influence missing data can have on the final conclusion. Based, on these considerations, it was concluded that there is plausible evidence that Cr (VI), administer via drinking water, acts via a mutagenic MOA for carcinogenicity (Akerman *et al.*, 2009).

References

Akerman, G., N. McCarroll, J. Chen, N. Keshava, A. Kligerman and E. Rinde (2009), *An Evaluation of the Mode of Action Framework for Mutagenic Carcinogens: Chromium (VI)*, Society of Toxicology, Baltimore, MD.

McCarroll, N., N. Keshava, J. Chen, G. Akerman, A. Kligerman and E. Rinde (2009), "An Evaluation of the Mode of Action Framework for Mutagenic Carcinogens Case Study II: Chromium (VI)", *Environ Mol Mutag*, 51(2):89-111.

McCarroll, N., N. Keshava, M. Cimino, M. Chu, K. Dearfiled, C. Keshava, A. Kligerman, R. Owen, A. Protzel, R. Putzrath, R. Schoney, (2008), "An Evaluation of the Mode of Action Framework for Mutagenic Carcinogens Case Study: Cyclophosphamide", *Environ Mol Mutag,* 49:117-131.

5. Example of liver mitogenic mode of action: conazole group of antifungal agents

Many of the conazole agents induce liver tumours in male mice. The key events leading to tumour induction are known: activation of constitutive receptors (CAR), modulation of cytochrome P450 CYP isoforms, increased liver weights, increased liver hypertrophy, induction of cell proliferation, and liver tumours. By expanding the sampling times and adding additional measurements such as liver enzymes (e.g., PROD, EROD, BROD and possible other liver function assays) and cell proliferation tests, the mode of action analysis can proceed and generate meaningful results long before the data from chronic studies are generated. Similarly, the organ weight data and histopathologic data on hypertrophy and hyperplasia can be examined to determine if a time-related increase in liver weight occurs and if the development of hyperplasia is sustained. It is also important to note that data pertinent to the MOA analysis may also be found in other toxicology tests such as other subchronic, chronic, carcinogenicity, developmental and/or reproductive toxicity studies. This is particularly true for target organ support, organ weights and histopathology and consistency across animal species.

References

Allen, J.W., D.C. Wolf, M.H. George, S.D. Hester, G. Sun, S.F. Thai, D.A. Delker, T. Moore, C. Jones, G. Nelson, *et al.* (2006), "Toxicity Profiles in Mice Treated with Hepatotumorigenic and Non-

hepatotumorigenic Triazole Conazole Fungicides: Propiconazole, Triadimefon, and Myclobutanil", *Toxicol. Pathol.*, 34, 853–862.

Wolf, D.C., J.W. Allen, M.H. George, S.D. Hester, G. Sun, T. Moore, S.F. Thai, D.A. Delker, E. Winkfield, S. Leavitt, *et al.* (2006),. "Toxicity Profiles in Rats Treated with Tumorigenic and Non-tumorigenic Triazole Conazole Fungicides: Propiconazole, Triadimefon, and Myclobutanil", *Toxicol. Pathol.,* 34, 895–902.

Nesnow, S., W. Ward, T. Moore, H. Ren and S.D. Hester (2009), "Discrimination of Tumorigenic Triazole Conazoles from Phenobarbital by Transcriptional Analyses of Mouse Liver Gene Expression", *Toxicol. Sci.*, 110, 68-83.

Appendix II. Example of Human Relevance Framework

Evidence Evaluation of the Human Health Relevance of Thiamethoxam-Related Mouse Liver Tumours (Pastoor *et al.*, 2005)

Thiamethoxam was shown to increase the incidence of mouse liver tumours in an 18-month study; however, thiamethoxam was not hepatocarcinogenic in rats. Thiamethoxam is not genotoxic, and given the late life generation of mouse liver tumours, suggests a time related progression of key hepatic events that leads to the tumours. These key events were identified in a series of studies of up to 50 weeks that showed the time dependent evolution of relatively mild liver dysfunction within 10 weeks of dosing, followed by frank signs of hepatotoxicity after 20 weeks leading to cellular attrition and regenerative hyperplasia. A metabolite CGA330050 was identified as generating the mild hepatic toxicity, and another metabolite, CGA265307, exacerbated the initial toxicity by inhibiting inducible nitric oxide synthase. This combination of metabolite generating hepatotoxicity and increase in cell replication rates is postulated as the mode of action for thiamethoxam. The relevance of these mouse specific tumours to human health was assessed by using the framework and decision logic developed by ILSI/RSI. The postulated mode of action was tested against the Hill criteria and found to fulfil the comprehensive requirements of strength, consistency, specificity, temporality, dose response, and the collective criteria of being a plausible mode of action that fits with known and similar modes of action. Whereas the postulated mode of action could theoretically operate in human liver, quantitation of key metabolites *in vivo* and *in vitro* showed that mice, but not rats or humans, generate sufficient amounts of these metabolites to initiate the hepatic toxicity and consequent tumours. Indeed rats fed 3000 ppm thiamethoxam for a lifetime did not develop hepatotoxicity or tumours. In conclusion, the coherence and extent of the database clearly demonstrates the mode of action for mouse liver tumourigenesis and also allows for the conclusion that thiamethoxam does not pose a carcinogenic risk to humans because of toxicokinetic differences between mice and humans.

3. STUDY DESIGN

3.1 DOSE SELECTION

3.1.1 Introduction

70. The purpose of a long-term bioassay (chronic toxicity and/or carcinogenicity studies) is the detection of biological evidence of any toxic and/or carcinogenic potential of the substance being investigated. Protocols should therefore maximise the sensitivity of the test without significantly altering the accuracy and interpretability of the biological data obtained. The dose regimen has an extremely important bearing on these two critical elements. Since one of the objectives is determination of the dose–response relationship in respect to any endpoints, the OECD TGs 451 (Carcinogenicity Studies), 452 (Chronic Toxicity Studies) and 453 (Combined Chronic Toxicity/Carcinogenicity Studies) normally require at least three dose levels, as well as controls.

71. OECD TGs 451, 452 and 453 outline general principles for dose selection in their respective bioassays. Provision of in depth guidance and a strategy for dose selection is however beyond the scope of the Test Guideline texts. This section of the Guidance Document is designed to underpin and expand the principles of dose selection for chronic toxicity and carcinogenicity studies outlined in the Test Guidelines.

72. These principles of dose selection are generally applicable to a wide range of chemicals, whatever their field of application e.g., pesticides, industrial chemicals and pharmaceuticals. However, although this document provides a number of references to specific requirements for dose selection for pharmaceuticals, the principles applied in studies on pharmaceuticals may differ from that for other agents (Rhomberg *et al.*, 2007; ICH, 2008). More information is generally available on the pharmacodynamic effects of pharmaceuticals, including the results of controlled clinical studies, than for other types of chemicals. The intended systemic human exposure is known and detailed pharmacokinetic studies enable valid comparisons to be made between the systemic exposures in rodents at the chosen dose levels and those in humans under therapeutic administration of the drug, as measured by the comparative areas under the curve (AUC) of blood concentrations over time (Rhomberg *et al.*, 2007). Users of the Guidance should therefore consult the Guideline S1C (R2) on dose selection for carcinogenicity studies of pharmaceuticals for specific information on testing of such chemicals (ICH, 2008).

73. General principles and guidance on dose selection for chronic toxicity and carcinogenicity studies in rodents are provided in two publications of the International Life Sciences Institute (ILSI). An initial 1997 report, entitled *Principles for the Selection of Doses in Chronic Rodent Bioassays* (ILSI, 1997), presented common views on the selection of doses for carcinogenicity and chronic toxicity studies while a second ILSI working group publication in 2007, entitled *Issues in the Design and Interpretation of Chronic Toxicity and Carcinogenicity Studies in Rodents: Approaches to Dose Selection* (Rhomberg *et al.*, 2007) provides additional discussion of the factors that influence dose selection in long-term bioassays (Rhomberg *et al.*, 2007). The latter publication incorporates concepts included in other documents prepared by national and international organisations (OECD, ECETOC, NTP and USEPA), and places emphasis on the influence of the objectives of a long-term bioassay on dose selection, as summarised in section 3.1.3 of this guidance. Users of this Guidance Document are recommended to consult these publications for more information on the factors influencing dose

selection. The following sections provide guidance on (a) the principles for dose selection in the Test Guidelines 451, 452, 453, and (b) the influence of the objectives of a long-term bioassay on dose selection.

74. The basic principles for the conduct of chronic toxicity and carcinogenicity studies will differ, given that the endpoints are different. However, given the drive to reduce the number of animals for welfare reasons and the cost of carcinogenicity bioassays, there is a need to maximise the results to assess non-cancer effects that may arise during the study, as these may be critical to the interpretation of any carcinogenic effects. The possibilities for considering non cancer effects in the interpretation of carcinogenic effects are maximised in the TG 453, Combined Chronic Toxicity/Carcinogenicity Study.

75. While the guidance provided in this chapter can be taken as generally applicable to dose selection for either a chronic toxicity or a carcinogenicity bioassay, or a combined chronic and carcinogenicity study, users of the guidance should be mindful of the primary objectives of the study in establishing the optimum study design in terms of dose selection.

76. In selecting appropriate dose levels for long-term bioassays (e.g., TG 451, TG 452, TG 453), a balance has to be achieved between hazard identification/characterization on the one hand and characterization of low-dose responses and their relevance on the other. This is particularly relevant in the situation where a combined chronic toxicity and carcinogenicity study (TG 453) is to be carried out.

3.1.2 Principles for Dose Selection

77. The general principles for dose selection laid down in the TGs are summarised as follows:

- Dose levels should generally be based on the results of shorter-term repeated dose or range finding studies and should take into account any existing toxicological and toxicokinetic data available for the test substance or related materials (Barton at al. 2006).

- The highest dose level should be chosen to identify toxic effects including the principal target organs while avoiding severe toxicity, morbidity, or death (OECD 2000, GD No.19). It should be noted that the severity of toxicity and survival in a two year study may be underestimated from the short-term study; for this reason, Test Guidelines indicate that a top dose lower than the dose providing evidence of toxicity in a short-term study may be chosen. When there is no toxicity in shorter-term studies it is recommended to consult with the relevant regulatory authorities.

- Dose levels should be selected to reflect the purpose of the study. In most cases, dose levels and dose level spacing may be selected to establish a dose-response and to derive a point of departure (e.g., BMDL or NOAEL).

78. These principles for dose selection are broadly similar to the key principles for dose selection outlined in the ILSI publications (ILSI, 1997; Rhomberg et al., 2007), as listed in full in Appendix I to this section. They are further discussed in the following sections.

3.1.2.1 Key information for the selection of doses in chronic toxicity and carcinogenicity studies

79. Identifying/characterizing carcinogenic effect is the primary objective of the OECD TG 451 on Carcinogenicity Studies while identifying/characterizing other toxic effects is the primary objective of the OECD TG 452 on Chronic Toxicity Studies. The OECD TG 453 Combined Chronic Toxicity/Carcinogenicity Studies combines the objective of OECD TG 451 and OECD TG 452. For all three studies the core minimum study design comprises at least one control group and three dose groups, each of which is exposed to different concentrations of the test substance.

80. The robustness of a carcinogenicity or chronic toxicity study, in particular the former, is dependent on a demonstration that the dose levels selected in the study are adequate to show an effect or effects of the test substance, without producing either false negative results (because the doses selected were too low) or false positive results (because metabolic/homeostatic mechanisms are overwhelmed, etc), which may be problematic in assessing risk in humans.

81. The data provided by shorter-term repeated dose or range finding studies, including 28-day or 90-day studies, are important in selecting the dose levels for a longer-term chronic toxicity or carcinogenicity study. The dose levels used in such studies and the NOAELs established can be used as a starting point for dose selection, both in relation to the highest dose level to be chosen in the study and possibly (but not necessarily) to the lower dose levels. Considerations that should be taken into account in determining whether similar, lower or higher dose levels than those used in a short-term study should be selected for a chronic toxicity or carcinogenicity study include (Rhomberg *et al.*, 2007):

- whether the effect is an adaptive response (e.g., liver hypertrophy in the absence of any other evidence of hepatotoxicity);

- potential of the toxic effect(s) observed in repeat dose toxicity studies of shorter duration to progress to neoplasia. A dose that induces a marked effect in such study should not be excluded from a carcinogenicity study if the effect or effects can reasonably be anticipated to be a precursor event in the development of neoplasia (e.g., a key event for the mode of action of the test substance). However, care should be taken that selection of a dose level that induces such effects will not result in excessive toxicity in the carcinogenicity study;

- the potential that an effect may limit the sensitivity of the chronic/carcinogenicity study to detect tumours (e.g., haemolytic anemia may limit the duration of the study due to an increase in mortality or to severe toxicity that may compromise the health of the animals; for more examples see Appendix 1, *Principle 5*, ILSI Principles for Dose Selection in Chronic Rodent Bioassays);

- the duration of the short-term study (e.g., repeated dose 90-day study, repeated dose 28-day study, two-generation reproduction study) and the potential for a toxic effect to progress in severity (e.g., progression from focal to multifocal necrosis);

- evidence of transitory effects that may be life-threatening: if prechronic studies revealed transitory effects that may last during some days or weeks until metabolic capacity (e.g., by liver enzyme induction) is adapted, testing of high dose is limited by transitory effects of life-threatening nature (e.g., sedation);

- use of gavage for administration of the test substance in studies of shorter duration. A dose that induces overt toxicity in a gavage study may be tolerated if a dietary route of

administration is selected for a carcinogenicity study because of the differences in toxicokinetics and toxicodynamics resulting from the two methods of administration.

82. Additional evidence on the extent to which dose levels should be increased or decreased in a long-term study relative to a short-term study or studies may be provided by dose–response data from the latter studies. For example, a marked reduction in dose levels would be warranted if results from short-term studies show that a minor increase in dose is associated with a pronounced increase in severity or incidence of a lesion (i.e., a steep dose–response). It is recommended that all the information from such short-term studies, (rather than the use of an arbitrary factor e.g., one-tenth the highest dose tested in a short-term study that induced a severe toxic effect) should be used when selecting the high dose (or mid and low dose levels) for a proposed carcinogenicity study.

83. Available toxicokinetic data (ADME) should always be taken into account when selecting dose levels for a chronic toxicity or carcinogenicity study, although such data may not be readily available for all chemicals, as they are not required under all regulatory schemes. Many toxicokinetic processes influencing absorption, distribution, elimination and metabolic activation or detoxication may become saturated at higher doses, resulting in systemic exposures to parent compound or metabolites that would not be expected in the real life human exposures for which risk assessments are needed. The effect of repeated exposures on the pattern of absorption, metabolism, detoxification, and clearance of a compound will provide information on the internal dose achieved during chronic exposure under conditions of the bioassay. The importance of having data on toxicokinetics in reaching a decision on the design most suitable for a chronic toxicity or carcinogenicity study is stressed in this guidance and the use of such data are discussed in more detail in Chapter 3, Section 3.4 of this Guidance Document.

84. Physiologically-based toxicokinetic (PBTK) modelling is also a valuable tool for defining doses where non-linear toxicokinetics may occur, thus allowing this to be considered in selecting the highest and other dose levels in the study. The use of PBTK modelling is explored in more detail in Section 3.4. Finally, specific mechanistic studies (where available) may provide useful information regarding target tissues affected by the test substance and the doses associated with effects on key events, and should be taken into account when selecting doses for a chronic toxicity or carcinogenicity study.

85. Additional considerations in selecting dose levels for chronic toxicity or carcinogenicity studies arise as a result of practical constraints such as the physicochemical characteristics of the substance to be tested (e.g., solubility, vapour pressure), palatability of the compound in food or drinking water, and other factors such as the potential for the test substance to cause adverse effects such as irritancy at the site of administration (Rhomberg *et al.*, 2007). Further guidance on these aspects is provided in the ILSI publications (ILSI, 1997; Rhomberg *et al.*, 2007) and in Chapter 3, Section 3.5 of this Guidance Document.

86. Information on, and consideration of, the mode of action (MOA) of a suspected carcinogen is particularly important, since the dose selection may differ depending on the known or suspected mode of action (Sonnich-Mullin *et al.*, 2002; Cohen *et al.*, 2003; Meek *et al.*, 2003; Holsapple *et al.*, 2006; Boobis *et al.*, 2006; EPA 2005). In selecting dose levels for such a study, doses will need to be placed carefully, to yield observations of subtle precursor effects or other biomarkers of toxicity without inducing confounding effects related to frank toxicity. This approach requires some previously gathered information on potential modes of action, e.g., from genotoxicity studies. Further guidance on these aspects is discussed in Chapter 2 of this Guidance Document.

3.1.2.2 Selection of the top dose

87. Dose selection should be based on the findings of subchronic or range-finding studies. The highest dose level to be used in a chronic toxicity or carcinogenicity study needs to be carefully considered and the reasons for the final choice clearly defined. Ideally, the dose levels selected will maximise the detection of dose–response relationships and facilitate the extrapolation of these to potential hazards for other species, including humans.

88. The selection of the highest dose level to be used in a chronic toxicity or carcinogenicity study has long been a matter of controversy. At the time when long-term animal bioassays began to be routinely used to assess the qualitative potential of a test substance to cause chronic toxicity and cancer, the emphasis was on testing at high levels in order to maximise the potential of such studies to detect effects. The concept of the Maximum Tolerated Dose (MTD), conventionally defined as the highest dose to produce toxic effects without causing death and to decrease body weight gain by no more than 10% relative to controls (OECD 2002, GD No. 35) became well established. The MTD is often used in the assessment of a chronic toxicity or a carcinogenicity study to decide whether the top dose tested was adequate to give confidence in a negative result. This Guidance Document focuses on the selection of the top dose, rather than attempting to define an MTD.

89. While some regulatory bodies or organisations interpret an adequate high dose to be a minimally toxic dose, others emphasize the need to select a dose level that is a maximally tolerated dose (i.e., more severe toxicity should be demonstrated). Thus, because of differences in views regarding the severity of toxic effects that are interpreted as providing evidence that an adequate high dose has been attained or exceeded, a completed carcinogenicity bioassay may be considered to be acceptable by one organisation but not by another. Many carcinogenicity studies can be challenged on the basis of selection of a top dose that is too high, particularly if there is a large interval to the next highest dose. This results in data that are difficult to interpret and cannot be used for regulatory purposes. Appendix 2 of Rhomberg (Rhomberg *et al*., 2007) provides detailed guidance on criteria that can be applied in order to assess the acceptability of the high dose level or MTD.

90. If the main objective of the study is to identify a cancer hazard, there is broad acceptance that the top dose should ideally provide some signs of toxicity such as slight depression of body weight gain (not more than 10%), without causing e.g., tissue necrosis or metabolic saturation and without substantially altering normal life span due to effects other than tumours. Excessive toxicity at the top dose level (or any other dose level) may compromise the usefulness of the study and/or quality of data generated. Criteria that have evolved for the selection of an adequate top dose level include: (in particular) toxicokinetics; saturation of absorption; results of previous repeated dose toxicity studies; the MOA and the MTD.

91. Toxicokinetic non-linearity should also be considered in the selection of the top dose to be used. Although top dose selection based on identification of inflection points in toxicokinetic non-linearity may result in study designs that fail to identify traditional target organ or body weight effects, it must be appreciated that metabolic saturation in fact represents an equivalent indicator of biological stress. In this case, the stress is evidenced by appearance of non-linear toxicokinetics rather than appearance of histological damage, adverse changes in clinical chemistry, haematology parameters or decrease in body weight gain (Toxicokinetics is discussed in Section 3.4).

92. For compounds that are (or potentially are) genotoxic, conventional considerations of top dose given above would apply. For compounds that are not genotoxic, the top dose should be informed by considerations of MOA (see Chapter 2). For a given compound, for which the mode of action is known or suspected, establishing a point of departure based on precursor key events, may also be protective against any carcinogenic effect. This is because non-genotoxic carcinogens produce

cancer by perturbing normal physiology or biochemistry. The long-term assay should be designed to also identify and characterize these key events.

93. Nutritional effects, physiological factors, physical-chemical factors and compound bioavailability can influence selection of the top dose level to be used in a long-term bioassay. For nutritional and possibly other physiological reasons a maximum level is imposed, commonly 5% concentration in the diet (Sontag *et al.*, 1976; Chhabra *et al*., 1990).

94. Palatability of a compound in either feed or water can also lead to perturbation of physiological homeostasis or nutritional status. A compound's solubility limit or vapour pressure may constrain selection of the top dose level. Irritation at the site of compound deposition may constrain dose or otherwise confound cross species extrapolation. Inhalation of doses that overwhelm pulmonary clearance may lead to tissue responses that are specific to the species being tested; however, this does not apply to asbestos-like substances. These limitations may influence selection of the top dose.

95. The top dose used in the study may be based on a defined level of the target population's exposure of interest and multiples of that exposure (e.g., 100 times or 150 times higher based on dose ratios expressed in terms of body weight). If toxicokinetic data are available, dose levels based on internalised doses (e.g., AUC) can be used. It has been shown that the relative systemic exposure corresponds (AUC ratios) better with dose ratios expressed in terms of body surface area rather than ratios expressed in terms of body weight. It should be noted that the use of systemic exposure comparison between rodents and humans to derive the top dose may be useful in the case of pharmaceuticals testing (ICH, 3(R2), S1C(R2)) but is not likely to be useful for testing plant protection products or commodity chemicals. Given the uncertainties regarding exposure levels in scenarios where these are used, and given the need for these chemicals that the top dose be sufficiently high for the purpose of classification for carcinogenic effect, an inherent potential of the substance in question to cause cancer irrespective of the dose should be demonstrated.

96. The relevance of the top dose level recommended to be used in the study to potential human exposures can also be debated. Mechanistic information gleaned from this type of study may be irrelevant. Positive results in the high dose group may be difficult to interpret as they may reflect a high-dose-only phenomenon, not relevant for human exposure. On the other hand, if the top dose level is set lower, to ensure relevance, the power of the study to detect effects may be compromised.

3.1.2.3 Dose level spacing

97. Selection of dose intervals is influenced by the study objectives (see section 3.1.3) and the available information. Dose levels and dose level spacing may be selected to establish a dose-response and to derive a point of departure (e.g., BMDL or NOAEL). The dose level spacing does not need to be regular. It may be reduced in regions of the dose-response curve where particularly robust estimation is needed, e.g., in the range of the anticipated BMD or a suspected threshold. The increasing emphasis on consideration where the lower dose levels used in the study are placed, and the number of such dose levels, reflects the changing purposes of lifetime bioassays.

98. If the primary purpose is identification of hazard, whether this is chronic toxicity or carcinogenicity, the focus of dose selection should be on maximizing the power of the study and on the top doses tested. As the risk assessment process becomes increasingly concerned with characterization of human risk, there has been a corresponding need to characterize whether and how high- dose effects extend to responses at lower exposure levels as well, with a consequent interest in how the lower dose levels are placed in bioassays (Rhomberg *et al.*, 2007).

99. Dose selection and dose level spacing need to be based, where possible, on the following considerations:

- known or suspected non-linearities or inflection points in the dose–response;
- toxicokinetics, and dose ranges where metabolic induction, saturation, or non-linearity between external and internal doses does or does not occur;
- precursor lesions, markers of effect, or indicators of the operation of key underlying biological processes;
- key (or suspected) aspects of mode of action, such as doses at which cytotoxicity begins to arise, hormone levels are perturbed, homeostatic mechanisms are overwhelmed, etc.;
- regions of the dose–response curve where particularly robust estimation is required, e.g., in the neighbourhood of the anticipated point of departure;
- consideration of anticipated human exposure level;
- a suspected threshold.

100. Dose levels should be selected to reflect the purposes of the study, and they should use available knowledge on how dose-dependent biological and impacted physiological factors may affect study outcomes. The Test Guidelines (TG 451, paragraph 24; TG 452, paragraph 24; and TG 453, paragraph 26) indicate that "The dose level spacing selected will depend on the characteristics of the test substance, and cannot be prescribed in this Guideline, but two to four fold intervals frequently provide good test performance when used for setting the descending dose levels and addition of a fourth test group is often preferable to using very large intervals (e.g., more than a factor of about 6-10) between dosages. In general, the use of factors greater than 10 should be avoided, and must be justified if used".

101. If prior evidence allows, it may be possible to optimise the design in terms of the location of the dose levels. A design often applied uses a mid dose that is half of the top dose, or the geometric mean of the low and high dose. This will ensure that the power and sensitivity of the assay is maximised and that at least one dose is unlikely to have a carcinogenic or other effect. This approach minimises the chance of a false negative (failing to detect an effect that actually exists) at some increased risk of a false positive (finding a high-dose effect that is an artefact of excessively high doses and is not relevant to the dose range of interest).

102. Limited information may be obtained regarding the shape of the dose–response curve, particularly if non-linearity is seen in the middle of the dose range. The power of the assay at lower dose levels will also be limited if the incidences of the responses of interest are low (e.g., rare tumours) and not markedly different from the controls. Information on the dose-response relationship would depend on how well the dose range of interest is anticipated.

103. The issue of where to place the lowest dose should receive comparable attention as to the placement of the top dose. If the lowest dose is too low, it may be insufficiently powerful and therefore uninformative; if too high, it may lose opportunities to characterize effects as near as possible to environmental exposure levels. For example, for pharmaceuticals, a dose sufficient to produce a pharmacodynamic effect or result in systemic exposure comparable with that expected at the intended clinical use is normally selected for the low dose level (see also ICH guideline S1C(R2)). For agrochemical products, food additives and similar products, selection of the lowest dose/dietary concentration usually takes into account a "desired" NOAEL reflecting the likely use of this NOAEL for risk assessment purposes by application of an Uncertainty Factor (UF) to obtain an Acceptable Daily Intake (ADI).

104. It may be possible to place adjacent dose levels somewhat above and below the levels at which a key transition in underlying biological actions, including considerations of the mode of action, is believed to lie, thereby revealing its influence on response. Transitions need not be sharp; typically, there are ranges of doses over which an underlying biological factor, such as metabolic saturation or cytotoxicity, comes increasingly into play.

105. When evaluating threshold effects in a chronic toxicity bioassay, the doses selected will include at least one dose high enough to show toxicity, at least one dose low enough to show lack of toxicity, and usually one but occasionally more than one in between to help characterize the shape of the curve near the point where the threshold appears to lie (Rhomberg *et al.*, 2007). These dose placement concerns differ from those in the carcinogenicity bioassay for substances where genotoxicity is known or suspected; however, this difference disappears if the BMD approach is used. The same dose range is preferred for both phases of a combined chronic toxicity/carcinogenic study, particularly if the MOA is under investigation.

3.1.3 Integration of the objectives of a long-term bioassay

106. The ILSI publications (ILSI, 1997; Rhomberg *et al.*, 2007) provide practical guidance on factors that influence dose selection in long-term bioassays, with particular emphasis on how the varying objectives of a chronic toxicity/carcinogenicity bioassay influence dose level selection.

107. Test Guideline 453 (combined chronic toxicity/carcinogenicity studies) identifies nine possible objectives:

- The identification of the carcinogenic properties of a chemical, resulting in an increased incidence of neoplasms, increased proportion of malignant neoplasms or a reduction in the time to appearance of neoplasms, compared with concurrent control groups;
- The identification of the time to appearance of neoplasms;
- The identification of the chronic toxicity of a chemical;
- The identification of target organ(s) of chronic toxicity and carcinogenicity;
- Characterization of the dose:response relationship;
- Identification of a point of departure (e.g., BMDL or NOAEL);
- Extrapolation of carcinogenic effects to low dose human exposure levels;
- Prediction of chronic toxicity effects at human exposure levels;
- Provision of data to test hypotheses regarding mode of action (EPA 2005; OECD 2009, GD No.116; Boobis *et al.*, 2006; Cohen *et al.*, 2003; Holsapple *et al.*, 2006; Meek *et al.*, 2003).

108. Various study designs have been proposed to address these objectives, as described by Rhomberg *et al.*, 2007. The core study design for a long-term bioassay as laid out in TGs 451, 452, 453 primarily addresses the objective of identification/characterization of carcinogenic substances or those causing other toxic effects, while seeking to integrate the other objectives as far as possible. Modifications of the core study design in TGs 451, 452, 453 in order to optimise data for the other objectives may compromise the Mutual Acceptance of Data and should be discussed with the relevant regulatory authorities before the commencement of the study. More generally, it is recommended to

ensure regulatory acceptance of the study design before performing any long-term bioassay to be submitted to a competent authority.

109. The different objectives outlined above seek to maximise the statistical power of the study at different points on the dose-response curve. The focus may be on a level of response or on the shape and slope of the overall curve. The situation is also complicated by the fact that, below a certain dose, attempts to increase statistical power by increasing animal numbers in particular dose groups become futile.

110. For the majority of bioassays, there will be one primary objective (typically the identification of carcinogenic potential and/or chronic toxicity) and several subsidiary objectives such as characterizing the dose-response curve, extrapolating to low doses, or identifying a point of departure. The nature of the subsidiary objectives will be contingent on the intended outcome. If a valid negative result is obtained in a carcinogenicity study (e.g., OECD TG 451), and this was the only objective of the study, there may be no further questions to be answered. If a positive result is obtained, however, a number of issues arise regarding the nature of the carcinogenic responses and their relevance to the levels of exposure of target populations, requiring further investigation into the nature and interpretation of the effects seen. Consideration of the mode of action framework before embarking on a long-term bioassay will provide guidance on optimising the design to collect the information necessary to the interpretation (see Chapter 2).

111. The choice of dose levels for the identification of a point of departure will depend on the type of point of departure sought. For a NOAEL, a dose without effects is required but ideally at the highest dose at which this can be observed. For a BMDL, the data from all dose levels is used but it is important that the responses differ at the different doses.

112. The study design selected at the outset should include dose levels that combine several objectives. One approach to achieve this is to include additional dose groups in such a way that the optimal doses for a number of different objectives are all included (Rhomberg *et al.*, 2007). Some doses would be optimised for some objectives and others for other objectives, essentially running several bioassays in tandem.

113. However, this is not feasible, given animal welfare, economic and time constraints. When attempting to combine these various objectives into a single study, selection of dose levels must be done in a way that does not compromise the primary objective while still allowing a secondary objective to be pursued in an acceptable albeit suboptimal manner. There may be embellishments to the core design based on study objectives but it would be a rare event when an erosion of the core minimum would be acceptable.

114. Based on Rhomberg *et al.*, 2007, four core selection schemes are presented below:

1. **Hazard Identification/Characterization Plus Dose–Response:** The top dose is chosen to increase the study's statistical power to detect effects that may be rare. A second dose combines two functions: (1) hedging against the top dose being found to have been too high in retrospect, and (2) providing the opportunity for dose–response characterization of any effects found. The lowest dose level or, if considered necessary, other lower dose levels can be placed so as to inform dose–response, no-effect levels, or other purposes. Key challenges will be balancing statistical power and toxicological relevance of the top dose level and compromising among subsidiary objectives while accounting for relevant dose-related physiological changes when setting lower dose levels.

2. **NOAEL/BMDL-Seeking for Threshold Effects:** The main aim is to identify no-effect (or low-effect) levels for the more sensitive adverse threshold effects. The top dose should aim at

engendering an adverse effect, the lowest dose should aim at constituting a NOAEL/BMDL, and the intermediate dose or doses should be set so as to identify the dose levels at which the top dose responses begin to manifest.

3. **Assessment of Safety of Human Exposure Levels:** This is modeled on safety assessment studies for nutrients and pharmaceuticals. For agents that are not genotoxic, show low toxicity, and evince no known difference in metabolic profile between rodents and humans, one can test multiples of anticipated human exposure. Lack of adverse effects at doses sufficiently above human exposure (and the perceived implausibility of non-threshold effects) gives evidence supporting the safety of the anticipated exposures. The bioassay exposures should be selected on an appropriate basis for animal:human comparison; for instance, the application to pharmaceuticals is typically based on area under the blood concentration-time curve that results from anticipated human exposures.

4. **Special-Purpose Bioassays**: this case would be beyond the OECD Test Guidelines and therefore is not developed in this document.

REFERENCES

Barton, H.A., T.P Pastoor, K. Baetcke, J.E. Chambers, J. Diliberto, N.G. Doerrer, J.H. Driver, C.E. Hastings, S. Iyengar, R. Krieger, B. Stahl and C. Timchalk (2006), "The Acquisition and Application of Absorption, Distribution, Metabolism, and Excretion (ADME) Data in Agricultural Chemical Safety Assessments", *Crit Rev Toxicol,* Vol. 36: 9-35.

Boobis, A.R., S.M. Cohen, V. Dellarco, D. McGregor, M.E. Meek, C. Vickers, D. Willcocks and W. Farland (2006), "IPCS Framework for Analyzing the Relevance of a Cancer Mode of Action for Humans", *Crit Rev in Toxicol*, 36:793-801.

Chhabra, R.S., J.E. Huff, B.S. Schwetz and J. Selkirk (1990), "An Overview of Prechronic and Chronic Toxicity/Carcinogenicity Experimental Study Designs and Criteria Used by the National Toxicology Program", *Environmental Health Perspectives,* 86: 313-321.

Cohen, S.M., M.E. Meek, J.E. Klaunig, D.E. Patton and P.A. Fenner-Crisp (2003), "The Human Relevance of Information on Carcinogenic Modes of Action: An Overview", *Crit Rev Toxicol,* 33:581-589.

ECETOC (1996), *Practical concepts for dose selection in chronic toxicity and carcinogenicity studies in rodents*, ECETOC Monograph No. 25. European Centre for Ecotoxicology and Toxicology of Chemicals. Brussels, Belgium.

EPA (2005), *Guidelines for Carcinogen Risk Assessment,* Risk Assessment Forum U.S. Environmental Protection Agency, Washington, DC
http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=116283&CFID=1267360&CFTOKEN=65052793&jsessionid=9830b2c4116e3d8fbbf017414e1a782e7f79TR

Holsapple, M.P., H.C. Pitot, S.N. Cohen, A.R. Boobis, J.E. Klaunig, T. Pastoor, V.L. Dellarco and Y.P. Dragan (2006), "Mode of Action in Relevance of Rodent Liver Tumors to Human Cancer Risk", *Toxicol Sci,* 89:51-56.

ICH (International Conference on Harmonisation) (2008), *Guidance for Industry. S1C(R2) Dose Selection for Carcinogenicity Studies,* U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER), September 2008 ICH**.** Revision 1

ICH (International Conference on Harmonisation) (2009), *ICH Harmonised Tripartite Guideline - Guidance on Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorisation for Pharmaceuticals – M3(R2)*
Available at: http://www.ich.org/cache/html/501-272-1.html

ILSI (International Life Sciences Institute) (1997), *Principles for the Selection of Doses in Chronic Rodent Bioassays*, Foran JA (Ed.). ILSI Press, Washington, DC.

Meek, E.M., J.R. Bucher, S.M. Cohen, V. Dellarco, R.N. Hill, L.D. Lehman-McKemmon, D.G. Longfellow, T. Pastoor, J. Seed and D.E. Patton (2003), "A Framework for Human Relevance analysis of Information on Carcinogenic Modes of Action", *Crit Rev Toxicol,* 33:591-653.

OECD (2000), *Guidance Document on the Recognition, Assessment, and use of clinical signs as humane endpoints for experimental animals used in safety evaluation,* OECD Series on Testing and Assessment No. 19, OECD, Paris.

OECD (2002), *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies*, Series on Testing and Assessment No. 35, OECD, Paris. Available at:

http://www.olis.oecd.org/olis/2002doc.nsf/LinkTo/NT00002BE2/$FILE/JT00130828.PDF.

OECD (2009), *Carcinogenicity Studies*, OECD Guideline for the Testing of Chemicals, TG No. 451, OECD, Paris.

OECD (2009), *Chronic Toxicity Studies,* OECD Guideline for the Testing of Chemicals, TG No. 452, OECD, Paris.

OECD (2009), *Combined Chronic Toxicity\Carcinogenicity Studies*. OECD Guideline for the Testing of Chemicals, TG No. 453, OECD, Paris.

Rhomberg, L., K. Baetcke, J. Blancato, J. Bus, S. Cohen, R. Conolly, R. Dixit, J. Doe, K. Ekelman, P. Fenner-Crisp, P. Harvey, D. Hattis, A. Jacobs, D. Jacobson-Kram, T. Lewandowski, R. Liteplo, O. Pelkonen, J. Rice, D. Somers, A. Turturro, W. West and S. Olin (2007), "Issues in the Design and Interpretation of Chronic Toxicity and Carcinogenicity Studies in Rodents: Approaches to Dose Selection", *Crit Rev Toxicol*, 37:729-837.

Smith, G. (2002), "Scientific Review of Chronic Rodent Bioassays. Criteria for Maximum Tolerated Dose (MTD) Attainment: Clinical Pathology- and Pathology-related Endpoints", Unpublished draft, submitted to Pest Management Regulatory Agency, Canada, September 18, 2002.

Sonnich-Mullin, C., R. Fielder, J. Wiltse, K. Baetcke, J. Dempsey, P. Fenner-Crisp, D. Grant, M. Hartley, A. Knaap, D. Kroese, I. Mangelsdorf, E. Meek, J. Rice and M. Younes (2001), "IPCS Conceptual Framework for Evaluating a Mode of Action for Chemical Carcinogenesis", *Reg Toxicol & Pharm,* 34:146-152.

Sontag, J.M., N.P. Page and U. Saffiotti (1976), *Guidelines for Carcinogen Bioassay in Small Rodents*, U.S. Dept. of Health and human services publication (NIH 76-801), National Cancer Institute, Bethesda.

Appendix 1: ILSI Principles for Dose Selection in Chronic Rodent Bioassays[2] (Rhomberg* et al., *2007)

Principle 1

Dose selection for chronic studies must be based on sound toxicologic principles. Within a reasonable dose range, increasing the dose can increase the ability to detect an effect; therefore, doses for chronic rodent bioassays should be selected within this range to maximize the sensitivity of a chronic bioassay. However, trying to increase study sensitivity by increasing doses into ranges that do not reflect application of sound toxicologic principles could lead to results that are inappropriate for human risk assessment.

Increasing the highest dose in a chronic bioassay may increase sensitivity within some defined dose range, but the potential exists that different mechanisms of toxicity or chemical mode of action are active at higher doses, which may not be relevant to humans exposed to lower doses. In this case, selection of the highest dose may be influenced by consideration of the mechanism/ mode of action and other factors discussed in Principle 4. However, when the highest dose in a carcinogenicity assay is limited by effects (e.g., a mode of action in one organ system) that are thought not to occur in humans, one must be aware that it still is possible that a higher dose of the chemical may be carcinogenic in other animal/organ systems.

To address these issues, the ILSI working group encourages an approach to dose selection that incorporates all relevant information from prechronic studies and other sources, uses toxicologic tools associated with an understanding of the mechanisms or mode of action by which a chemical produces an effect (e.g., genotoxicity, cell proliferation, etc.), and uses good scientific principles to enhance the accuracy of judgments of potential human risks. In the case of negative studies (particularly where the highest dose is chosen based on a full characterization of the chemical's toxicity in prechronic studies), use of sound scientific principles as well as all available chemical, physical, and toxicologic data will lessen concern that the result may be a false negative. Similarly, in positive carcinogenicity studies, this approach will lessen concern that the result may be a false positive. In both cases, the predictiveness of the bioassay for human health effects will be improved.

Principle 2

A chronic bioassay requires a major investment in resources and time, and the objective of such a study should be broader than hazard identification. Scientists who conduct chronic bioassays and those who use data from bioassays, including regulatory agencies, should encourage innovative approaches to dose selection by considering appropriate study designs, mechanistic data, and other information in the design and interpretation of studies. Use of additional endpoints and other information must be based on sound scientific rationale, and such designs should be evaluated on the basis of their individual merits.

A goal of high-dose selection in carcinogenicity bioassays is, in the context of hazard identification, to reduce the likelihood of a false-negative result. However, it is recognized that the qualitative nature of the hazard (e.g., carcinogenic response) may itself be dose dependent. This principle encourages approaches to dose selection that incorporate consideration of mechanistic and other toxicologic information. Such approaches should improve the scientific basis for dose selection and aid in interpretation of data generated from chronic bioassays.

---

[2] The terminology used in the Guidance Document regarding "long-term bioassay" designing both chronic toxicity and carcinogenicity studies doesn't apply to this annex.

Principle 3

Human exposure should be considered in dose selection, particularly for selection of the middle and lowest doses. Further, the middle and lowest doses should be selected to characterize the shape of the dose response curve as much as possible. Selection of the middle and lower doses should take into account factors such as the mechanism or mode of action, toxicokinetics, and others listed in Principles 4 and 5 and should not be based solely on a fraction of the highest dose.

Issues that should be considered when incorporating potential human exposure in dose selection include the human exposure route and mode, the dose range in the chronic bioassay in relation to human exposure, and the duration and frequency of human exposure, if known. Subpopulations that may be more highly exposed than the general population, or that are genetically more susceptible, also should be considered. The relationship between external and delivered (internal) dose (e.g., ingested dose versus dose delivered to the target organ, toxicokinetics) in both humans and test organisms may influence dose selection. Further, for substances expected to exhibit a toxicity threshold, or if the evaluation of carcinogenic potential is being combined with an evaluation of chronic toxicity, the study should be designed to include one dose that does not elicit adverse effects; that is, one dose should be a NOAEL. Of course, caution must be exercised to ensure that the NOAEL is not simply an artifact of small sample size or poor study design.

Principle 4

The [ILSI] working group has recommended the use of innovative approaches, additional endpoints, and other information in the selection of doses for chronic rodent bioassays. The following endpoints, generally determined in prechronic studies, should be considered in dose selection for chronic rodent bioassays. Further, it is recognized that endpoints other than those listed below may provide important information for dose selection, and use of those endpoints, where they are based on sound toxicologic principles, is encouraged. Such endpoints may be available presently, or they may be developed as the science of toxicology advances.

- **Histopathology**

  The site, morphology, and severity of the treatment-related effects observed in the pre-chronic study should be taken into account in setting dose levels for the chronic study. Histopathological examination of tissues, especially the liver, kidneys, gastrointestinal tract, urinary tract, respiratory tract, skin, spleen/bone marrow/blood, and endocrine tissues derived from properly designed pre-chronic studies, often provides information that is crucial for dose selection in chronic bioassays.

- **Toxicokinetics**

  Studies to determine the effect of dose (or exposure concentration) on absorption, tissue distribution, metabolism, and clearance of a compound are helpful in selecting appropriate doses for the chronic bioassay. The kinetics of absorption will determine the internal exposure dose achieved. The absorption and clearance of the compound and its metabolites will determine the systemic and target organ exposure resulting from a single dose and can be used to design the treatment regimen required to achieve a desired internal dose. The effect of repeated exposures on the pattern of absorption, metabolism, biotransformation, and clearance of a compound will provide information on the internal dose achieved during chronic exposure under conditions of the bioassay. The nutritional status of the chronically exposed animals may be affected during the experimental

period which is why adequate information on interactions between the exposure chemical and nutrionally important compounds may be of great value in the interpretation of the final results of the chronic study.

- **Cell Proliferation**

In the process of chemical carcinogenesis, events related to induced cell proliferation might be critical in fixing mutations and in providing a selective growth advantage to pre-cancerous cells. Considerations may be different for direct mitogenic stimulation of organ growth versus regenerative cell proliferation, and these modes of action should be distinguished for the test agent. Further, apoptosis can be a strong determinant of normal and pre-cancerous cell turnover kinetics and should be considered. Information on the dose dependence of regenerative cell proliferation is a useful adjunct to histological observations in determining the shapes of organ-specific toxic response curves. This information, when available, can be of value in selecting high, middle, and low doses and in interpreting the results of the study.

- **Physiological Functions**

Disturbances of physiology or homeostasis that would compromise the validity of the study should be considered in the dose-selection process. Examples include hypotension, inhibition of blood clotting, overwhelming normal pulmonary clearance mechanisms, immune system effects, and in some cases hormonal imbalance. Such disturbances, and their effects on the validity of a study, may be difficult to determine and may apply differentially to different categories of chemicals (e.g., pharmaceuticals in which the desired pharmacological action is a physiological effect).

- **Body Weight**

It is suggested that body-weight changes are the primary factor in the selection of the highest-dose group (that is, when no other toxic effects are observed), a decrement in body-weight gain of no more than 5–10% in pre-chronic studies should be used in the selection of the highest dose for chronic assays of carcinogenicity.

Historically, scientists have adopted a 10% decreased body-weight gain at the end of pre-chronic studies (typically 90 days duration) as the target that should not be exceeded in chronic (carcinogenicity) studies. It is now recognised that there is a positive correlation between body weight and the occurrence of certain tumours in rodent species and strains used in safety assessment or for hazard identification; i.e., the higher the body weight between 6 and 18 months on test, the higher the probability that the animal will develop some tumours. Moreover, the lower the body weight, the less sensitive the animal may be to agent-induced toxicity, including cancer. A significant decrease in body-weight gain therefore could reduce the animal's ability to respond to compound-induced toxicities.

- **Clinical Chemistry, Haematology, Urinalysis**

Clinical chemistry and urinalysis results are best used to support dose-selection decisions based on other criteria/parameters. Changes in serum clinical chemistry in the absence of histopathological observations may not affect high-dose selection but may complement dose-selection decisions based on toxicokinetics, cell proliferation, and other parameters. Haematology results also are more often affected secondarily to other processes more relevant for dose selection (e.g., inflammation). However, when haematological tissues are determined to be a target organ in pre-chronic studies, haematology results may be an appropriate basis for dose selection.

- **Organ Weights**

Organ weights are not often the critical factor in selection of doses for chronic rodent bioassays. Chemically induced changes in organ weights, however, should be considered in conjunction with other data in the dose-selection process.

Ideally, data from the factors and endpoints listed above would be collected from pre-chronic studies and used to select doses for chronic studies; however, not all parameters may be useful or necessary for every compound. Even when based on information concerning the points described above, dose selection for a chronic bioassay will remain an inexact process; thus, reconsideration of these same points must be made

when interpreting and assessing the significance of the effects obtained in the bioassay.

Principle 5

Physicochemical factors (e.g., solubility, vapour pressure), the bioavailability of the compound, the palatability of the compound in food or drinking water, and other factors such as the potential for the substance to cause adverse effects at the site of administration (e.g., irritation, erosion, and ulceration) will influence the selection of the highest dose for chronic rodent bioassays. It is recommended that doses for chronic rodent bioassays be selected to minimise or avoid adverse nutritional, physical, organoleptic, and irritant effects.

3.2 ROUTES OF EXPOSURE AND DOSE ADMINISTRATION CONSIDERATIONS

Introduction

115. The choice of the route of administration depends on the physical and chemical characteristics of the test substance, its intended field of application, the availability of information on shorter-term repeat dose studies and the predominant route of human exposure. Comparison across the set of studies available for a given chemical will be more robust if the parameters of the study e.g., route, dose volume, vehicle, remain constant between all the repeat dose toxicity studies. The three main routes of administration used in chronic toxicity and carcinogenicity studies are oral, dermal and inhalation. For example, if human exposure to the test substance is likely to be through food or is a pharmaceutical intended to be taken by mouth, the relevant route of administration will be the oral route, while for a workplace gas, inhalable dust or volatile liquid, inhalation should be the route of choice. The dermal route may be chosen, e.g., for substances used in the workplace, where skin contact is likely, or for pharmaceuticals applied to the skin. Other routes such as subcutaneous or intraperitoneal injection have been used when they are considered to be more appropriate for the anticipated route of exposure of humans (see paragraphs 136 and 137). In choice of route of administration of the test chemical due consideration should be given to animal welfare including choice of routes different from the predominant route of human exposure, if relevant.

116. Given the potential for oral exposure to a wide range of chemicals and also the practical experimental considerations associated with the long duration of chronic toxicity and carcinogenicity studies, the oral route is the route most commonly used in chronic toxicity and carcinogenicity studies. Route-to-route extrapolation may be considered for systemic effects when reliable data on ADME are available, rather than carrying out an additional study by a second route. For example, it may be possible to carry out an assessment of systemic effects via inhalation exposure based on the results of an oral chronic toxicity or carcinogenicity study (Gerrity and Henry, 1990). The use of route-to-route extrapolation should be decided on a case-by-case basis (Nielsen *et al.*, 2008) and is not, however, relevant for the assessment of local toxicity.

3.2.1 The oral route of exposure

117. Test substances may be administered via the diet or drinking water, by oral administration in capsules (in non-rodents) or by gavage, normally in a vehicle, depending on the physical and chemical characteristics of the test substance, its intended field of application and the predominant oral route of exposure of humans. Each method has advantages and disadvantages, and it should in particular be kept in mind that the toxicokinetics of the test substance may be affected by the method of oral administration. Data from previous short-term toxicological studies, including data on toxicokinetics, can provide information on potential local gastrointestinal effects and the extent of bioavailability of the test substance, in order to select the most appropriate route of oral administration and to demonstrate that systemic exposure is adequate (see also Section 3.4). As indicated in the Test Guidelines, a top dose not exceeding 1000 mg/kg body weight/day may apply except when human exposure indicates the need for a higher dose level to be used.

118. The animals are dosed with the test substance daily (seven days per week), normally for the entire duration of the study. Any other dosing regime, e.g., five days per week, needs to be justified. In the case of rodents, dosing of the animals should begin as soon as possible after weaning and acclimatisation and preferably before the animals are 8 weeks old.

3.2.1.1 Administration via the diet

119. Oral administration via the diet is the preferred route of administration if human exposure to the test substance is also likely to be via the diet. This route of administration may be appropriate if the objective is to establish an Acceptable Daily Intake (ADI) or Tolerable Daily Intake (TDI), for example for substances deliberately added to food, substances released from food contact materials or for environmental contaminants entering the food chain, and the pattern of exposure is continuous ingestion of small doses. However oral gavage studies may also be used to derive an ADI or a TDI.

120. When using oral administration via the diet, the test substance is administered in the diet either as a constant dietary concentration (mg/kg diet), or as a constant dose level in terms of the animal's body weight. In the latter case the dietary concentration must be adjusted regularly based on anticipated food consumption and body weight of the animals. While doses are expressed in terms of mg/kg diet, food consumption must be monitored on a cage basis at least weekly in order to be able to derive the intake of the test substance on mg/kg body weight per day or mg/m^2 per day. The food intake e.g., in the rat decreases from above 100 g per kg bw per day in early life (6-8 weeks, at the commencement of the study) to about little above 50 g per kg bw per day for older females (e.g., 6 months or more) and below 50 g per kg bw per day for older males. This will lead to a gradual decrease in intake of dietary administered test substance over age when keeping the dietary concentration of the test chemical constant. The concentration of the chemical in the feed should not normally exceed an upper limit of 5% of the total diet (FDA, 1982; Borzelleca, 1992), although higher levels are feasible (e.g., when testing carbohydrates or proteins) as long as the diet is adapted nutritionally adequately, e.g., the test substance is incorporated, at the expense of other components in a purified diet (Howlett *et al*., 2003).

121. Oral administration via the diet has the advantage that no handling of the animals is required. However, the palatability of the diet may be reduced at high dietary levels due to the taste or odour of the test substance, resulting in reduced food intake and thus reduced exposure to the test substance. This is likely to have been identified in previous shorter-tem studies and may require the introduction into the study design of an additional control group, pair fed (i.e., having matched food intake) in parallel with the high dietary level test group (see section 3.5.2 for further details). The substance should be stable during the preparation, storage and period of administration of the diet, for example it should not react chemically with dietary constituents, and analytical data must be provided to demonstrate this. It is also essential to ensure that the substance is mixed homogeneously in test diet at the desired level and, again, analytical data must be provided to demonstrate this, as required under Good Laboratory Practice (OECD, 1998).

3.2.1.2 Administration via drinking water

122. Oral administration in drinking water is the method of choice if human exposure to the test substance is likely to be via drinking water (e.g., drinking water contaminants) or in liquids (e.g., for substances that are volatile, or reactive with feed components, or any case where drinking water has an advantage over diet administration such as for soft drinks or beverages). The test substance is normally incorporated at a fixed concentration in the drinking water, at the approximate levels (in mg/ml water) required to provide the dose levels selected for the study (in mg/kg body weight per day), based on anticipated water consumption of the animals. Suspensions of test chemicals have been used where a solution in water has not provided a sufficiently high dose. Care should be taken to prevent precipitation of the test substance, with consequent accumulation in the drinking water valve, resulting in a disproportionately high dose being administered to the animal. Realistically, the test material should be soluble in water at all concentrations tested. While doses are expressed in terms of mg/ml water, water consumption must be monitored on a cage basis at least weekly in order to be able

to derive the intake of the test substance on mg/kg body weight per day. Concerning possible adjustment of the concentration of the test substance in the drinking water e.g., when this route of administration causes changes in water consumption due to palatibility of the drinking water, similar measures as described for dietary dosing are appropriate (see paragraphs 120 and 121) (Sharp and Regina, 1998; Wolfensohn and Lloyd, 1998; Pool, 1999; Nielsen *et al.,* 2008). The test substance should not markedly affect the palatability of the drinking water or cause marked changes in the pH, and its content and stability must be demonstrated analytically, as required under GLP (OECD, 1998).

3.2.1.3 Administration via gavage or encapsulation

123. Oral intubation (gavage) may be used if administration in the diet or drinking water is not possible, e.g., because of stability or palatability considerations. However, in the interests of animal welfare in particular, administration of the test substance by oral gavage should preferentially be restricted to those agents for which a bolus dose administration reasonably represents potential human exposure (e.g., administration of pharmaceuticals or food supplements orally at one or more doses per day) (Craig and Elliott, 1999; Brown *et al.*, 2000). Gavage dosing is experimentally more difficult than dietary administration, inducing stress in the animals which has toxicological implications (Brown *et al.*, 2000), and also requires daily handling of the animals, which may interfere with experimental parameters e.g., if neurobehavioural assessments are carried out during the study. A second control group to address this potential confounding factor is, however, rarely required and should be justified.

124. If the test substance is administered by gavage, this should be done using a stomach tube or a suitable intubation cannula, at similar times each day. The test substance may be administered in capsules, dissolved or suspended in a suitable vehicle. Administration by encapsulation rather than gavage dosing is a common route for dogs, but is not a preferred route for rodents, due to the associated technical difficulties. Vehicles of choice include aqueous solutions of thickeners such as methyl cellulose or carboxymethylcellulose, although other vehicles may be used (Gad *et al.,* 2006). Methylcellulose (0.5-1.0%) may have advantages over carboxymethylcellulose due to its superior wetting properties and polyethylene glycol (e.g., PEG 400) may also be used. Preferred vehicles are those that do not induce effects in their own right in long-term studies. Oil (e.g., corn oil) has been used when it is not possible to prepare aqueous solutions or homogeneous suspensions. When oil is used as a vehicle for gavage administration of the test substance dietary adjustment, for example a low fat diet should be considered, to compensate for the additional caloric intake. Induction of pancreatic adenomas in F-344 rats has been reported after administration of corn oil alone by gavage, with male F-344 rats being more sensitive than females (Boorman *et al.,* 1987; Haseman & Rao 1992). The maximum volume of solution that can be given by gavage in one dose depends on the size of the test animal (Diehl *et al.,* 2001; Gad *et al.,* 2006). For rodents, the volume ordinarily should not exceed 0.5-1 ml/100 g body weight, except in the case of aqueous solutions where 2 ml/100g body weight may be used (Diehl *et al.,* 2001). It should be noted, however that dosing volumes above 1 ml/100 g body weight may result in reflux of the dose.

125. Normally a single dose will be administered once daily, but where, for example, a substance is a local irritant or the pattern of human dosing is multiple doses per day, the daily dose may be administered as a split dose e.g., twice a day, within a 6 hour period. Variability in dose volume should be minimised by adjusting the concentration to ensure a constant volume at all dose levels. Potentially corrosive or irritant substances may however need to be diluted to avoid severe local effects, and testing at concentrations that are likely to be corrosive or irritant to the gastrointestinal tract should be avoided. The frequency and length of time for which the animals in a chronic toxicity or carcinogenicity study are dosed can lead to irritation in the oesophageal tissue and distress of the animals, potentially compromising the integrity of the study. If oral gavage is used, careful

observation should be conducted after dosing to watch for signs of distress such as laboured breathing, sudden lethargy, or poor mucous membrane colour.

3.2.2 The dermal route of exposure

126. The dermal route of exposure has been used in long-term carcinogenicity studies, primarily in the assessment of carcinogens such as polycyclic aromatic hydrocarbons in skin painting studies generally carried out in the mouse. Assessment of systemic toxicity or carcinogenicity using the dermal route is only appropriate if it has been demonstrated that the test substance is bioavailable via the skin, i.e., it crosses the skin barrier and the tested concentration of the substance shows either no or minimal irritation potential. Although the dermal route may be used in assessing the chronic toxicity and carcinogenicity of substances such as workplace chemicals, where skin contact is likely, or for pharmaceuticals applied to the skin, for which continuous dermal contact is anticipated, it is in practice a difficult route for long-term administration of a test substance, and the animal welfare implications involved in carrying out a 1- or 2-year dermal exposure study should be considered.

127. Bioavailability by the dermal route may be assessed initially via a dermal penetration study to determine extent of absorption through the skin *in vivo* or *in vitro* in accordance with OECD TG 427 or OECD TG 428 (OECD 2004a, 2004b) and following the guidance laid down in GD 28, the OECD Guidance Document for the conduct of skin absorption studies (OECD 2004c). This should be followed by a short term 21/28 day dermal toxicity study (TG 410, OECD, 1981a) or a subchronic dermal toxicity study (TG 411, OECD, 1981b) before conducting the longer term study. If the chemical does not penetrate the dermal layer, it is not appropriate to use this route for examination of chronic toxicity.

128. The method is based on the repeated application of the test substance, generally at a defined concentration in mg/ml in a suitable vehicle, to a clipped or shaved area of skin of approximately 10% of the total body surface area, to provide the desired dose in mg/kg body weight per day. Application is for at least 6 hours per day, 7 days per week, for a period of 24 months. Animals are normally housed separately in dermal studies to prevent grooming behaviors and oral ingestion of the test substance. TG 410 on Repeat Dose Dermal Toxicity: 21/28 day study (OECD, 1981a) or TG 411: Subchronic Dermal Toxicity: 90 day study (OECD, 1981b) should be consulted in the case of testing carried out by the dermal route,

129. The site may be occluded with polyethylene sheeting and gauze patches or semi-occluded, in order to prevent dislodgement of material and oral ingestion, which could affect the validity or usefulness of the study. With volatile or semi-volatile materials, application and covering procedures should minimise the possibility of evaporation. However, long-term dermal studies without occlusion are acceptable when justified scientifically, considering the potential difficulties of occlusion in long-term studies, including stress in the animals and the resource demands of repeated occlusion. In particular long-term dermal studies in the mouse without occlusion are recommended, considering the technical difficulties of occlusion when using this species.

3.2.3 The inhalation route of exposure

130. If it is likely that humans may be exposed by inhalation to a test substance, either as a gas, a vapour, or a liquid or solid aerosol (or a mixture thereof), the inhalation route should be used to evaluate the chronic toxicity or carcinogenicity of the substance in animals. The results of acute, subacute (28 days), subchronic (90 days) and range finding inhalation studies should be considered

when designing these studies and selecting concentration levels that will yield robust data regarding local and systemic toxicity. When testing substances which are irritants and/or corrosive, existing information should be used in selecting an appropriate dilution ratio for testing (see paragraph 133). This guidance does not extend to the testing of nanoparticles which can pose challenging physiological and methodological problems.

131. A chronic inhalation toxicity or carcinogenicity study should follow the principles described in Test Guideline 413 (subchronic inhalation toxicity: 90 day study, OECD, 2009a) in all respects except for the number of animals per group and study duration. Exposure by the inhalation route is normally carried out for 6 hours per day, 7 days per week, but exposure for 5 days per week may also be used, if justified. A rationale must be provided when using an exposure duration less than 6 hours per day. If rodent species smaller than rats, e.g., mice, are exposed nose-only, maximum exposure durations may be adjusted to minimise species-specific distress. The exposure restraining tubes must be adapted to the size of this species. Historical data, and/or published information in scientific literature may be used to demonstrate that the exposure technology chosen does not impose undue stress to the animals exposed. Endpoints suggestive of undue immobilization stress-related effects include marked changes body temperature, changes in ventilation, or decrease in body weight gains. These effects have been seen in animals exposed to clean air in nose-only exposure tubes for varying lengths (van Eijl *et al*., 2006; Narciso *et al.,* 2003). There is the potential for stress experienced by animals in nose-only exposure studies to affect the results of the study by affecting immune competence (Bernet *et al.,* 1998; Jakab and Hemenway 1989) or xenobiotic metabolism (Fechter *et al.,* 2008), or by modulating gene expression in key organs (Thomson *et al*., 2009; Ha *et al*., 2003; Laconi *et al.,* 2000; Panuganti *et al*., 2006; Sato *et al*., 2006; Yin *et al.,* 2006), although some did not find confounding effects (Rothenberg *et al.,* 2000). The duration of a chronic toxicity study by the inhalation route will normally be 12 months and that for a carcinogenicity study will be 18-24 months, dependent on the species used (see also Section 3.3.3 of this guidance). Further general guidance on the performance of an inhalation toxicity study can be found in the Guidance Document 39 on Acute Inhalation Toxicity Testing (OECD, 2009b) and the OECD Guidance Document 125 on Histopathology for Inhalation Toxicity Studies (OECD, 2010). Although GD 39 is intended to provide guidance for acute inhalation studies, the technical aspects of exposing animals and generating and characterizing test atmospheres are similar for repeated exposures and single exposures and are therefore also applicable for chronic and carcinogenicity inhalation studies.

132. The nature of the test substance and the object of the test should be considered when selecting an inhalation chamber. For studies of liquid or solid aerosols and for vapor that may condense to form aerosols, the nose-only exposure method allows the avoidance of oral exposure due to grooming of particles deposited on the fur. However, the welfare implications of a 1- or 2-year nose-only exposure study, and the potential for physiological effects of stress experienced by the animals to affect the results of the study, can lead to preference for the use of the whole-body mode of exposure (Thomson *et al.,* 2009). Reasons for choice of exposure system should be justified in the study report. Particular attention should be paid to the technical problems that may arise from the large numbers of animals in whole body inhalation chambers (e.g., time required to attain inhalation chamber steady-state, heat and CO_2 production, and adsorption of test article on inhalation chamber walls and other surfaces). To ensure atmosphere stability when using a whole-body chamber, the total volume of the test animals should not exceed 5% of the chamber volume, and there should be a sustained dynamic airflow of at least 10 air changes per hour. Principles of the nose-only and whole body exposure techniques and their particular advantages and disadvantages are addressed in GD 39.

133. Test substances that are irritating or corrosive should always be tested using methodology laid out in Test Guideline TG 413 because it provides the study director or principal investigator with control over the selection of target concentrations. Corrosive or irritating test substances should be tested at concentrations that will yield the desired degree of toxicity without affecting longevity or undue stress to respiratory tract irritation (GD 39; OECD, 2009b). Any information available on the

corrosive or irritancy potential of the substance, including existing *in vivo* and *in vitro* data, pH values, and data from similar substances, should be considered in the design of a long-term study. Results from existing inhalation studies can be used to determine appropriate dose levels and should be reviewed carefully before any new range finding studies are undertaken. When exposing animals to corrosive or irritating substances, the targeted concentrations should be low enough to not cause marked pain and distress, yet sufficient to extend the concentration-response curve to levels that reach the regulatory and scientific objective of the test. These concentrations should be selected on a case-by-case basis, preferably based upon adequately designed range-finding studies that provide information regarding the critical location of irritation within the respiratory tract and endpoint for probing it. Adequately designed range-finding studies should demonstrate whether respiratory tract irritation depends on any irritation threshold (concentration-dependent) or on the total daily exposure intensity (concentration x time – dependent), and whether carry-over effects from one exposure day to another may lead to time-dependent exacerbations. Some irritant effects are instant in onset and others require time to accumulate. These factors need to be identified and may serve as justification for dose selection.

134. Species selection should be carefully considered for test substances causing upper respiratory tract irritation because numerous secondary species-specific physiological responses make the extrapolation from small rodents to humans more difficult (GD 39, OECD, 2009b). In depth justification for species-selection is necessary when using species other than rats for inhalation studies of irritant test substances.

135. For substances likely to accumulate in the lung over time due to poor solubility or other properties, the degree of lung-overload and delay in clearance needs to be estimated based on adequately designed pre-studies; ideally a 90-day study with postexposure periods long enough to encompass at least one elimination half-time. The use of concentrations exceeding an elimination half-time of approximately 1 year due to lung-overload at the end of study is discouraged.

3.2.4 Other routes of exposure

136. Other routes of exposure e.g., subcutaneous or intraperitoneal injection are generally only used in chronic toxicity or carcinogenicity studies when they mirror the anticipated route of administration in humans, such as in the case of pharmaceuticals. For example subcutaneous or intramuscular injections may be used for pharmaceuticals and for materials designed to be used as implants or prostheses. The subcutaneous and intraperitoneal routes have also been used in carcinogenicity bioassays for some solid-state, insoluble materials (e.g., fibres and plastics).

137. For substances administered parenterally, the dose volume used, stability of the formulation before and after administration, pH, viscosity, osmolality, buffering capacity, sterility and biocompatibility of the formulation are factors to consider (Diehl *et al.,* 2001). The smallest needle size should be used for administration, taking into account the dose volume, viscosity of injection material, speed of injection and species. (Diehl *et al.*, 2001). The use of parenteral injections in chronic toxicity and carcinogenicity studies is likely to result in local inflammation, and has significant animal welfare implications. Study investigators should document compelling reasons for using this method of administration.

REFERENCES

Bernet, F., J.F. Dedieu, C. Laborie, V. Montel and J.P. Dupouy (1998), "Circulating Neuropeptide Y (NPY) and Catecholamines in Rat under Resting and Stress Conditions. Arguments for Extra-adrenal Origin of NPY, Adrenal and Extra-adrenal Sources of Catecholamines", *Neurosci Lett,* 250(1):45-8.

Boorman, G.A., D.A. Banas, S.L. Eutis and J.K. Haseman (1987), "Proliferative Exocrine Pancreatic Lesions in Rats. The Effect of Sample Size on Incidence of Lesions", *Toxicol Pathol,* 15: 451-456.

Borzelleca, J.F. (1992), "Macronutrient Substitutes: Safety Evaluation", *Regulatory Toxicology and Pharmacology*, 16, 253-264.

Brown, A.P., N. Dinger and B.S. Levine (2000), "Stress Produced by Gavage Administration in the Rat", *Contemporary Topics in Laboratory Animal Science,* 39(1): 17-21.

Craig, M.A. and J.F. Elliott (1999), "Mice Fed Radiolabeled Protein by Gavage Show Sporadic Passage of Large Quantities of Intact Material into the Blood, an Artifact not Associated with Voluntary Feeding", *Contemporary Topics in Laboratory Animal Science,* 38(5): 18-23. ISSN: 1060-0558.

FDA (1982), *Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives Used in Food*, US Food and Drug Administration, Washington, DC.

Derelanko, M.J. and M.A. Hollinger (Eds) (1995), *CRC Handbook of Toxicology*, CRC Press, Boca Raton, Florida.

Diehl, K.-H., R. Hull, D. Morton, R. Pfister, Y. Rabemampianina, D. Smith, J.-M. Vidal and C. van de Vorstenbosch (2001), "A Good Practice Guide to the Administration of Substances and Removal of Blood, Including Routes and Volumes", *Journal of Applied Toxicology*, 21:15-23. Available at: http://www.ff.up.pt/farmacologia/pdf/good_practice_lab_animals.pdf

Fechter, L.D., A. Nelson-Miller and C. Gearhart (2008), "Depletion of Liver Gluthathione Levels in Rats: a Potential Confound of Nose-only Inhalation", *Inhal Toxicol*, 20:885-890.

Gad, S.C., C.D. Cassidy, N. Aubert, B. Spainhour and H. Robbe (2006), "Nonclinical Vehicle Use in Studies by Multiple Routes in Multiple Species", *International Journal of Toxicology*, 25: 499-521.

Gerrity, T.R. and C.J. Henry (1990), *Principles of Route-to-route Extrapolation for Risk Assessment: Proceedings of the Workshops on Principles of Route-to-route Extrapolation for Risk Assessment*, Hilton Head, SC and Durham, NC. New York: Elsevier.

Ha, H.Y., K.S. Kim, Y.I. Yeom, J.K. Lee and P.L Han (2003), "Chronic Restraint Stress Massively Alters the Expression of Genes Important for Lipid Metabolism and Detoxification in Liver", *Toxicol Lett,* 146(1):49-63.

Haseman, J.K. and G.N. Rao (1992), "Effects of Corn Oil, Time-related Changes and Inter-laboratory Variability on Tumour Occurrence in Control Fischer 344 (F344/N) Rats", *Toxicol Pathol*, 20 (1): 52-60.

Hayes, A.W. (Ed) (1994), *Principles and Methods of Toxicology*, Third Edition, Raven Press, New York.

Howlett, J., D.G. Edwards, A. Cockburn, P. Hepburn, J. Kleiner, D. Knorr, G. Kozianowski, D. Müller, A. Peijnenburg, I. Perrin, M. Poulsen and R. Walker (2003), "The Safety Assessment of Novel Foods and Concepts to Determine their Safety in Use", *International Journal of Food Science and Nutrition*, 54 (Suppl.1), 1-32.

Jakab G.J. and D.R. Hemenway (1989), "Restraint of Animals Required for Nose-Only Inhalation Toxicologic Studies Suppresses Pulmonary Antibacterial Defenses", *Inhal Toxicol.* 1(4):289-300.

Laconi, E., C. Tomasi, F. Curreli, S. Diana, S. Laconi, G. Serra, M. Collu and P. Pani (2000), "Early Exposure to Restraint Stress Enhances Chemical Carcinogenesis in Rat Liver", *Cancer Lett,* 161(2):215-20.

Larson, J.L., D.C. Wolf and B.E. Butterworth (1994), "Induced Cytotoxicity and Cell Proliferation in the Hepatocarcinogenicity of Chloroform in Female B6C3F1 Mice: Comparison of Administration by Gavage in Corn Oil vs *ad libitum* in Drinking Water", *Fundam Appl Toxicol,* 22:90–102.

McClellan, R.O. and R.F. Henderson (Eds) (1989), *Concepts in Inhalation Toxicology*, Hemisphere Publishing Corp, New York.

Narciso, S.P., E. Nadziejko, L.C. Chen, T. Gordon and C. Nadziejko (2003), "Adaptation to Stress Induced by Restraining Rats and Mice in Nose-only Inhalation Holders", *Inhal Toxicol,* 15:1133-1143.

Nielsen, E., G. Østergaard and J.C. Larsen (2008), *Toxicological Risk Assessment of Chemicals*, Informa Healthcare, US, ISBN-13: 978-0-8493-7265-0.

OECD (1981a), *Repeated Dose Dermal Toxicity: 21/28-day Study*, OECD Guideline for the Testing of Chemicals, No. 410.

OECD (1981b), *Subchronic Dermal Toxicity: 90-day Study*, OECD Guideline for the Testing of Chemicals, No 411.

OECD (1998), *OECD Principles on Good Laboratory Practice,* Series on Principles of Good Laboratory Practice and Compliance Monitoring No 1, OECD, Paris.

OECD (2004a), *Skin absorption: in vivo method*, OECD Guideline for the Testing of Chemicals No 427, OECD, Paris.

OECD (2004b), *Skin absorption: in vitro method*, OECD Guideline for the Testing of Chemicals No 428, OECD, Paris.

OECD (2004c), *Guidance Document for the conduct of skin absorption studies,* Series on testing and assessment N°28, OECD, Paris.

OECD (2009a), *Subchronic Inhalation Toxicity: 90-day Study,* OECD Guideline for the Testing of Chemicals No 413, OECD, Paris.

OECD (2009b), *Guidance Document on Acute Inhalation Toxicity Testing,* Series on Testing and Assessment No 39, OECD, Paris.

OECD (2010), *Guidance Document on Histopathology for Inhalation Toxicity Studies, Supporting TG 41 (Subacute Inhalation Toxicity: 28-day Study) and TG 413 (Subchronic Inhalation Toxicity: 90-day Study)*, Series on testing and assessment No 125, OECD, Paris.

Panuganti, S.D., F.D. Khan and C.K. Svensson (2006), "Enhanced Xenobiotic-Induced Hepatotoxicity and Kupffer Cell Activation by Restraint-Induced Stress" *JPET*, 318:26–34.

Pool, T. (1999), *The UFAW Handbook of the Care and Management of Laboratory Animals,* Blackwell Science, Oxford UK.

Rothenberg, S., R.M. Parker, R.G. York, G.E. Dearlove, M.M. Martin, K.H. Denny, S.D. Lief, A.M. Hoverman and M.S Christian (2000), "Lack of Effects of Nose-only Inhalation Exposure on Testicular Toxicity in Male Rats", *Tox Sci*, 53:127-134.

Sato, T., H. Yamamoto, N. Sawada, K. Nashiki, M. Tsuji, K. Muto, H. Kume, H. Sasaki, H. Arai, T. Nikawa, Y. Taketani and E. Takeda (2006), "Restraint Stress Alters the Duodenal Expression of Genes Important for Lipid Metabolism in Rat", *Toxicology*, 227(3):248-61.

Sharp, P.E. and M.C.L. Regina (1998), *The Laboratory Rat,* Boca Raton FL USA: CRC Press.

Thomson, E.M., A. Williams, C.L. Yauk and R. Vincent (2009), "Impact of Nose-only Exposure System on Pulmonary Gene Expression", *Inhal Toxicol*, 21(S1): 74-82.

van Eijl, S., R. van Oorschot, B. Oliver, F.P. Nijkamp and N. Bloksma (2006), "Stress and Hypothermia in Mice in a Nose-only Cigarette Smoke Exposure System", *Inhal Toxicol,* 18:911-918.

Wolfensohn, S. and M. Lloyd (1998), *Handbook of Laboratory Animal Management and Welfare,* Blackwell Science, Oxford UK.

Yin, D., Y. Zhang, C. Stuart, J. Miao, Y. Zhang, C. Li, X. Zeng, G. Hanley, J. Moorman, Z. Yao and M. Woodruff (2006), "Chronic Restraint Stress Modulates Expression of Genes in Murine Spleen", *J Neuroimmunol,* 177(1-2):11-7.

3.3 CHOICE OF SPECIES AND STRAIN, NUMBERS AND SEX OF ANIMALS, STUDY DURATION, ALTERNATIVE *IN VIVO* MODELS

General issues

138. The choice of species to be used in a chronic toxicity or a carcinogenicity study is dictated by a number of factors, including the following:

- physiological and metabolic similarity to humans, in order to provide a valid model for extrapolation of the findings,
- familiarity with the species,
- availability of existing data on the species chosen,
- lifespan of the animals,
- ease of handling under experimental conditions,
- other issues such as cost of maintenance

139. Rodents (e.g., rats, mice or hamsters) have been used extensively, also dogs and primates. The choice of species must be justified and may be dictated by the purpose of the study (e.g., chronic toxicity or carcinogenicity) and by regulatory requirements. It should also be noted that mechanistic studies should be performed on the same species and strain as the cancer/chronic toxicity studies unless otherwise scientifically justified. With few exceptions, carcinogenicity and combined chronic toxicity/carcinogenicity studies are normally carried out in rodent species. Similarly, chronic toxicity studies are normally carried out in rodents. Chronic toxicity testing in non-rodents may however be required under certain regulatory regimes.

3.3.1 Testing in rodents

140. As indicated, rodent species have been used in the majority of chronic toxicity studies and in almost all carcinogenicity testing. The standard approach to carcinogenicity testing has been to use two species, the rat and the mouse, although in recent years a number of alternative approaches have been initiated to refine this approach in hazard identification and for risk assessment purposes, as discussed further in paragraphs 148 onwards. Advantages deriving from the use of rats and mice include the (relatively) low cost of maintenance, their short lifespan, meaning that a lifetime study can be completed in 2 – 3 years, and the availability of a large amount of historical data on age-related biochemical, haematological and pathological changes including on spontaneous tumours at specific organ sites.

141. Rodents have a number of metabolic pathways, physiological and pathological responses in common with humans. However, there are a number of instances where the chronic toxicity and carcinogenicity findings in rodents have been demonstrated not to be relevant to humans, because of toxicokinetic and toxicodynamic differences including species-specific pathways of metabolism, genetic differences, enzyme differences, differences in toxicologic pathways etc. If differences in toxicokinetics or toxicodynamics and/or other relevant parameters are suspected between the test species and humans that may have an impact on the relevance of the outcome of the study, these should be explored to determine if another test species may be more appropriate. Syrian golden hamsters have been used in some studies, for testing substances for which there is evidence that the toxicokinetics or toxicodynamics in humans are more similar to that in hamster than rats or mice. For example, hamsters have been used when considering a specific mode of action (PPAR-alpha, etc.).

Hamsters have also been used for studies of carcinogenesis in the respiratory and urinary tract, particularly for administration by parenteral routes such as intraperitoneal and intratracheal installation. It should be noted that a number of rodent tumour types are considered to be species-specific or strain-specific tumours with no relevance for humans (ECHA, 2009) (See Section 2.2).

142. It is important to consider the general sensitivity of the test animals, their background pathology and hence the responsiveness of particular organs and tissues to the chemicals under test when selecting rodent species, strains or stocks for toxicity studies. In general the selected rodent strain or stock should be well-characterized preferably including data on e.g., body and organ weight, haematological and biochemical parameters and background pathology. Additionally, it is important that test animals come from healthy colonies. Normally Specific Pathogen Free (SPF) animals are used, being SPF derived at birth and maintained under barrier conditions.

Rodent species and strain specificity

143. Assessment of chronic toxicity in rodents using the Test Guideline 452 is generally carried out in the rat, although other rodent species, e.g., the mouse, may be used. Since the normal lifespan of the rat is longer than that of the mouse, the potential for development of age-related background pathologies that may be influenced by the choice of a particular strain is less for rats than mice. In practice the strain commonly in use in the testing laboratory will be selected, since the laboratory will have historical data that will aid in the interpretation of any test substance-related change.

144. The Fischer 344 rat is a particularly well-characterized rat strain in carcinogenicity studies, since it has been the selected rat strain for the National Toxicology Programme (NTP) studies for over 20 years. However it has recently been reported (King-Herbert *et al.*, 2010) that the NTP is currently evaluating the Harlan Sprague Dawley (Hsd: Sprague Dawley SD) as the primary rat model for NTP studies, due to a number of health issues with the Fischer 344 rat and and decreased fecundity inherent in Wistar rats.

145. Importantly, in selecting a suitable rat strain for carcinogenicity testing, test animals should be selected that are likely to survive for the recommended duration of the study (see Section 3.3.2). Britton *et al.* (2004) reported that of the three rat strains studied (Harlan Hsd:Sprague-Dawley SD, Harlan Han Wistar Hsd:BrlHan:WIST, Charles River Crl:CD), Harlan Wistar strain survived in much greater numbers in 104-week carcinogenicity studies. The improved survival rate, according to the authors, appeared to be independent of body weight and food consumption and was reflected in the spontaneous pathology profile. Other authors believe this phenomenon to be attributable to a combination of obesity and genetic susceptibility and advocated dietary restriction as a method of extending survival in long-term carcinogenicity bioassays (Keenan, 1996). However, there is no scientific consensus on applying dietary restriction in such studies.

146. As discussed further in Section 3.5, and as reported by many investigators, dietary restriction results in a delay in age-related degenerative diseases such as nephropathy, which is commonly seen in all rat strains and has been shown to be diet-related. Dietary restriction may however result in a lower susceptibility of the animals to the development of tumours in carcinogenicity studies and to development of chemically-induced toxicity, thus presenting problems in extrapolation of the results of such studies to humans and, as indicated above, there is no scientific consensus on the application of dietary restriction.

147. Mouse strains used in carcinogenicity testing include the B6C3F1 mouse, as used by NTP, the ICR Swiss (CD-1), BALB/c, etc. The use of multiple strains of mice is being explored by NTP (King-Herbert *et al.*, 2010). The CD-1 mouse has been used by the US EPA OPP for chronic long-term toxicity studies. Notably, different mouse inbred strains show a variation in susceptibility to

tumourigenesis in different organs. The commonly used strains, in particular the B6C3F1 mouse used by NTP, carry hepatocellular tumour susceptibility loci that result in a high susceptibility to chemically induced hepatocarcinogenesis (Gariboldi *et al.,* 1993; Manenti *et al.,* 1994), which has limited their usefulness in carcinogenicity testing, while CD-1, an outbred mouse line derived from the Swiss strain has a relatively high incidence of spontaneous lung tumours and a high susceptibility to chemically induced lung tumourigenesis (Manenti *et al.,* 2003).

148. In recent years there has been considerable debate about the value of the two rodent species approach to carcinogenicity and about the continued use of the mouse as a second species, within the ICH (ICH, Proceedings of the Third International Conference, 1995) and in other fora (e.g., Huff and Haseman, 1991; Gold and Stone, 1993; Ennever *et al.*, 2003; Cohen, 2004, Billington *et al.*, 2010; Storer *et al.,* 2010). A number of studies have assessed the relative individual contribution of rat and mouse carcinogenicity studies and whether the use of rats or mice alone, or alternatively the reduced protocol using male rats and female mice would result in a significant loss of information on carcinogenicity relevant to human risk assessment. This debate has led to the suggestion that there may be no need for routine conduct of two long-term rodent carcinogenicity studies, since the use of the mouse in carcinogenicity testing may have limited utility (References above, also Griffiths *et al.,* 1994, Usui *et al.,* 1996, Carmichael *et al.,* 1997; Meyer, 2003, Doe *et al.,* 2006). However, testing in a second species is still acceptable, and is required under some current regulatory programmes.

149. Other experimental approaches to the evaluation of carcinogenic potential have been recommended, that may obviate the requirement to test in a second species (see also paragraphs 164 onwards). These approaches include short or medium-term *in vivo* rodent test systems providing insight into carcinogenic endpoints, such as models of initiation-promotion in rodents, or models of carcinogenesis using transgenic or neonatal rodents.

Numbers and sex of animals to be used in rodent studies

150. The Test Guidelines TG 451, TG 452 and TG 453 specify the core number of animals to be used in chronic toxicity and carcinogenicity studies. In a stand-alone chronic toxicity study, the core number indicated is normally at least 20 animals of each sex per group. Additional animals may also be included in the study design for interim kills during the study, also satellite animals for investigation of reversibility of any toxicological changes and sentinel animals for investigation of disease status. Smaller numbers of animals per sex and dose group are acceptable for these supplementary groups, as indicated in TG 452. It is unlikely that a regulatory authority would find a study using a lower core number of animals per sex and per group acceptable for regulatory purposes, unless a robust scientific justification is provided, since a sufficient number of animals should be used so that a thorough biological and statistical evaluation can be carried out. Furthermore, the Mutual Acceptance of Data (Council Decision C(81)30/FINAL – 12 May 1981, amended on 26 November 1997 - C(97)186/FINAL) would not apply to such a study, which may result in test duplication and many more animals being used. The key issues of importance in carrying out a statistical evaluation of the results of a chronic toxicity study are discussed further in Chapter 4 of this Guidance Document. In the case of a combined chronic toxicity and carcinogenicity study, there is provision for a smaller number of animals (at least 10 per sex and per group) to be used in the chronic toxicity phase, since interpretation of the data from the reduced number of animals per group in the chronic toxicity phase of this combined study will be supported by the data from the larger number of animals in the carcinogenicity phase of the study.

151. Similarly, TG 451 on the conduct of a carcinogenicity study specifies that at least 50 animals of each sex per dose group should be used, plus a concurrent control. Again, it is unlikely that a regulatory authority would find a study using a lower core number of animals per sex and per group acceptable for regulatory purposes, since a sufficient number of animals should be used so that a thorough biological and statistical evaluation can be carried out. It is however possible to increase

numbers of animals in all groups, in particular the lower dose groups, in order to increase the sensitivity of the study. In general use of additional numbers of animals above the 50 males and 50 females per group indicated in the TG for carcinogenicity testing (OECD TGs 451 and 453) has to be justified, considering e.g., animal strain, survival rate and statistical power. This is discussed further in Chapter 4. However, a number of publications have indicated that survivability problems exist for certain strains, notably the Sprague-Dawley rat (Nohynek *et al*., 1993; Keenan, 1996). For strains with poor survival such as Sprague Dawley rats, higher numbers of animals per group may be needed in order to maximise the duration of treatment (typically at least 65/sex/group).

3.3.2 Testing in non-rodents, including considerations of numbers of animals to be used

152. The use of non-rodent species may be considered when available data suggest that they are more relevant for the prediction of health effects in humans. Non-rodents are used generally only for chronic toxicity testing and not for carcinogenicity testing. The principles and procedures outlined in this Guidance Document, together with those outlined in OECD TG 409, Repeated Dose 90-day Oral Toxicity Study in Non-Rodents (OECD, 1998) should be applied, with appropriate modifications. The use of non-rodent species should be, in the main, restricted to special purpose studies, rather than for basic screening of chronic toxicity. As indicated in the Test Guideline, a second, non-rodent species should only be used:

- where effects observed in other studies indicate a need for clarification/characterization in a second species, or
- where toxicokinetic/toxicodynamic studies indicate that the use of a specific non-rodent species is the most relevant choice of laboratory animal, or
- where other specific reasons justify the use of a non-rodent species.

153. The dog has been a commonly used non-rodent species in chronic toxicity studies in the past. There has been extensive debate about the need for, and added-value of, chronic toxicity studies in the dog compared with a 90 days subchronic toxicity dog study (Box and Spielmann, 2005, Doe *et al.,* 2006, ESAC, 2006, EFSA 2007; Kobel *et al*., 2010). As a result of analyses carried out by these authors and also by the US EPA (Baetcke *et al.*, 2005), it has been suggested that tests of longer duration than 3 months using typical non-rodent species such as the dog do not have a substantial added value for making regulatory decisions. Some regulatory regimes have therefore now discontinued this requirement. For other regulatory sectors, such as small molecule pharmaceuticals, a large retrospective analysis confirmed the need for a 9 month study in the dog and provided examples for which even 6 months was not sufficient. For these reasons ICH (ICHM3R2, 2009) generally recommends 9-month studies in non rodents for chronic use pharmaceuticals.

154. Dogs used for chronic toxicity testing should be of a defined breed. Beagles are the most commonly used dog breed. The study design should minimise the numbers of animals used, and for a chronic toxicity study normally 4-6 animals per dose level are used. Dosing should begin preferably at four to six months and not later than nine months of age. Of particular importance when using dogs for toxicity testing, are considerations of appropriate housing, exercise, the need for environmental enrichment and for social contact. These aspects are discussed further in Section 3.5.

155. Other non-rodent species used include minipigs, as their basic physiology is considered to be very similar to humans, and they may therefore provide a better model than e.g., dogs or rodents. Rabbits, although used in the area of skin and eye irritation testing and reproductive toxicity testing, are rarely if ever used as a second species for chronic toxicity and carcinogenicity testing, and their use is therefore not discussed further in this Guidance Document.

156. Minipigs used for chronic toxicity testing should be of a defined breed. Gøttingen Minipigs are the most commonly used minipig strain. The study design should minimise the number of animals used; for a chronic toxicity study, normally 4-6 animals per dose level are used. Dosing should begin preferably at three to four months of age. Where the study is conducted as a preliminary to a long-term chronic toxicity study, the same species/breed should be used in both studies. Animal welfare considerations are of the utmost importance when using minipigs for toxicity testing, including housing, exercise, the need for environmental enrichment and for social contact. These aspects are discussed further in Section 3.5.

3.3.3 Study duration

157. The duration of the chronic toxicity study and of the chronic toxicity phase in the combined chronic toxicity/carcinogenicity study is normally 12 months, although longer or shorter periods may be used if scientifically justified, and for pharmaceuticals, chronic studies of 6 months duration in rats are required.

158. In the carcinogenicity study, mice are generally exposed to the test chemical for 18–24 months and rats for 24–30 months with exposure being longer for strains of greater longevity or with a lower spontaneous tumour rate. However, exposure for longer than 24 months is unusual and should be justified. TG 451 specifies that the duration of the study will normally be 24 months for rodents, representing the majority of the normal life span of the animals to be used. Shorter or longer study durations may be used, dependent on the lifespan of the strain of the animal species in the study, but should be justified. For specific strains of mice, e.g., AKR/J, C3H/J, CD-1 or C57BL/6J strains, for which documentation exists showing that a duration of 18 months may be more appropriate (e.g., Giknis and Clifford, 2010), a reference to this information is sufficient for the justification of using a duration shorter than 24 months. Many carcinogenicity studies in mice are conducted for 18 months; therefore, there is rather limited historical control data available at 24 months. The study may also make provision for interim kills, e.g., at 12 months, to provide information on progression of neoplastic changes and mechanistic information, if scientifically justified. Where such information is already available from previous repeat dose toxicity studies on the substance, interim kills may not be scientifically justified.

159. Termination of the study should be considered when the number of survivors in the lower dose groups or the control group falls below 25 per cent, considering the survival of each sex separately. The US EPA Health Effects Test Guidelines 870.4200 (US EPA, 1998b) specify that survival in any group should not fall below 50% at 15 months in the case of mice and 18 months in the case of rats, or below 25% at 18 and 24 months respectively. In addition, the WHO (1990) has recognised a further type of carcinogenicity study that continues until mortality in the most susceptible group reaches a fixed level, usually 80%. In addition the OECD GD 19 gives guidance on when to kill the animals in a study, based upon the recognition, assessment and use of clinical signs as humane endpoints (OECD 2000).

160. The study should not normally be extended beyond the point when the data available from the study are no longer sufficient to enable a statistically valid evaluation to be made. However, in the case where only the high dose group dies prematurely for obvious reasons of toxicity, this should not trigger termination. While the validity of the study may be prejudiced by early mortality, e.g., in the high dose group, valuable information will still be obtained from it, and a decision to terminate the study in its entirety must be carefully weighed against the animal welfare implications of having to repeat the study. The lower dose groups, in particular, the next highest dose level, when continued to the scheduled end of the study may still be useful for the evaluation.

161. If the current dosing regime results in severe animal toxicity and the study must be terminated before the full duration of exposure, the study sponsor needs to contact the regulatory authority immediately. All data should be compiled and all available tissues preserved for further evaluation. While this study may not meet all test guideline requirements for chronic/carcinogenicity testing, the results may be useful and considered in the overall risk assessment. The determination of a retest will be made on a case-by-case basis by the regulatory authorities.

Consideration of the acceptability of a negative carcinogenicity result relative to survival in the study.

162. For a negative result to be acceptable in a rat carcinogenicity bioassay, survival in the study should ideally be no less than 50% in all groups at 24 months, while for "life span studies", studies continued to end of life/ death of the animals survival at study termination should not be less than 25%. In a mouse study, survival in all groups in the study should be no less than 50% at 18 months. It is the responsibility of the study director to use rodent strains that would ensure adequate survival at 18/24 months. Additionally, no more than 10% of any group should be lost due to autolysis, cannibalism, or management problems.

163. Acceptability of a negative result in a carcinogenicity study based on survival rates will vary depending on the study design. For example, a survival rate of 50% may not be appropriate for a BMD design with appreciably more dose groups than the more conventional design. A flexible approach is necessary to accommodate different situations, most importantly the distinction between confidence in a negative result for risk characterization (most often required for risk assessments) and obtaining a quantitative dose-response value e.g., a BMD (most often required for risk/benefit analysis). Survival of less than 50% of animals in the top dose group need not disqualify the evaluation of a negative study outcome, provided that the higher mortality in this group can be clearly attributed to another toxic effect, such as chronic undernutrition or malabsorption resulting from gastrointestinal irritation by too high a dietary concentration of the test substance and no trend/drift/change in tumour incidence is observed. Evaluation of a negative study outcome may be based on calculation of the power of the test for groups with lower mortality.

3.3.4 Alternative in vivo *models for carcinogenicity testing, including testing in transgenic animals*

164. Some of the medium-term tests for carcinogenicity involve the development of proliferative lesions in a single tissue, e.g., foci of alteration in the liver (Williams *et al.*, 1982; Goldsworthy *et al.*, 1986; Ito *et al.*, 1989). Others use tumour end-points, such as induction of lung adenomas in the A-strain mouse (Maronpot *et al.*, 1986) or induction of tumours in initiation–promotion studies using various organs, including the skin, bladder, intestine, liver, lung, mammary gland and thyroid (see reviews by Enzmann *et al.*, 1998a & 1998b; IARC, 1992 & 1999). A further category of study is the "start/stop" protocol. Here, an agent is administered for a limited period to induce particular effects or lesions; the progression or reversibility of these is then observed in the absence of further treatment (Todd, 1986; Marsman & Popp, 1994).

165. Transgenic assays in genetically engineered rodents have also been developed following the identification of genes, such as proto-oncogenes and tumour-suppressor genes that are highly conserved across species and associated with a wide variety of human and animal cancers. The genetically engineered rodent designs involve activated oncogenes that are introduced (transgenic) or tumour suppressor genes that are deleted (knocked out). If appropriate genes are selected, these assay systems may provide supplementary information on mechanisms of tumour formation or serve as selective tests for carcinogens. The modified transgene is expected to accelerate carcinogen-induced cancer development without interfering with other relevant genetic and/or epigenetic steps. High

spontaneous tumour incidence in control animals is a major confounding factor of the conventional bioassay; the presence of the transgene itself does not induce high spontaneous tumour incidence in the short time span of the assay. These assays have been extensively reviewed in publications, including a single-theme issue of Toxicological Pathology (26 (4), 1998) and others (Tennant *et al.*, 1995; Contrera & DeGeorge, 1998; Eastin, 1998; Bucher, 1998; Eastin &Tennant, 1998; Santos *et al.*, 2008).

166. Transgenic mouse models are still under investigation for their utility in carcinogenicity testing, and have not yet been fully validated. Several of these models are, however, generally accepted as an alternative for 2-year mouse studies for pharmaceuticals. Transgenic mouse models may be useful as hazard identification / characterization screening models as part of an initial phase of the risk assessment process. However, they are not definitive proof of potential human carcinogenicity, and they are not proof of a specific mechanism of action. Like the 2-year bioassay, the results from tests in these models need to be incorporated into an overall integrated, weight of evidence evaluation for a given compound that takes into account genotoxicity, particularly DNA reactivity, structure activity relationships, results from other bioassays, and the results of other investigations including toxicokinetics, metabolism, and mechanistic information (ICH, 1997; Meyer, 2003; NAS, 2007; EFSA, 2009).

REFERENCES

Baetcke, K.P., W. Phang and V. Dellarco (2005), *A Comparison of the Results of Studies on Pesticides from 12- or 24-Month Dog Studies with Dog Studies of Shorter Duration,* US EPA Office of Pesticide Programs, Washington, DC.

Billington, R., R. Lewis, J. Mehta, I. Dewhurst (2010), "The Mouse Carcinogenicity Study is no Longer a Scientifically Justifiable Core Data Requirement for the Safety Assessment of Pesticides", *Critical Reviews in Toxicology,* Vol 40, 35-49.

Box, R.J. and H. Spielmann (2005), "Use of the Dog as Non-rodent Test Species in the Safety Testing Schedule Associated with the Registration of Crop and Plant Protection Products (Pesticides): Present Status", *Archives of Toxicology,* 79, 615-626.

Britton, L.D., N. Downes, P. Mullins and D. Mitchell (2004), "The Most Appropriate Strain for Rat Carcinogenicity Bioassay", *Toxicol Sci,* 78:560.

Bucher, J.R. (1998), "Update on National Toxicology Program (NTP) Assays with Genetically Altered or "Transgenic" Mice", *Environ Health Perspect,* 106: 619–621.

Carmichael, N.G., H. Enzmann, I. Pate and F. Waechter (1997), "The Significance of Mouse Liver Tumor Formation for Carcinogenic Risk Assessment: Results and Conclusions from a Survey of Ten Years of Testing by the Agrochemical Industry", *Environ Health Perspect*, 105:1196-1203.

Cohen, S.M. (2004), "Human Carcinogenic Risk Evaluation: an Alternative Approach to the Two-year Rodent Bioassay", *Toxicological Sciences*, 80:225-9.

Contrera, J.F. and J.J. DeGeorge (1998), "*In vivo* Transgenic Bioassays and Assessment of the Carcinogenic Potential of Pharmaceuticals", *Environ Health Perspect,* 106 (Sup. 1): 71–80.

Doe, J.E., A.R. Boobis, A. Blacker, V. Dellarco, N.G. Doerrer, C. Franklin, J.I. Goodman, J.M. Kronenberg, *et al*. (2006), "A Tiered Approach to Systemic Toxicity Testing for Agricultural Chemical Safety Assessment", *Critical Reviews in Toxicology,* 36, 37-68.

Eastin, W.C. and R.W. Tennant (1998), *Carcinogenicity: Testing Predicting & Interpreting Chemical Effects*, In: Kitchin K.T. (Ed), pp. 395–410, Marcel Dekker Inc., New York.

Eastin, W.C. (1998), "The U.S. National Toxicology Program Evaluation of Transgenic Mice as Predictive Models for Identifying Carcinogens", *Environ Health Perspect,* 106(Sup. 1): 81–84.

ECHA (2009), *Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging (CLP) of Substances and Mixtures*", Chapter 3.6, **"**Carcinogenicity", European Chemicals Agency.

EFSA (2007), *Opinion of the Scientific Panel on Plant Protection Products and their Residues on a request from the Commission related to the revision of Annexes II and III to Council Directive 91/414/EEC concerning the placing of plant protection products on the market - Toxicological and metabolism studies,* Question N° EFSA-Q-2006-118, EFSA Journal 449, 1-60.

EFSA (2009), *Existing Approaches Incorporating Replacement, Reduction and Refinement of Animal Testing: Applicability in Food and Feed Risk Assessment,* Question N° EFSA-Q-2005-231, EFSA Journal 1052, 1-77.

Ennever, F.K. and L.B. Lave (2003), "Implications of the Lack of Accuracy of the Lifetime Rodent Bioassay for Predicting Human Carcinogenicity", *Reg Tox Pharm,* 38, 52-7.

Enzmann, H., E. Bomhard, M. Iatropoulos, H.J. Ahr, G. Schlueter and G.M. Williams (1998a), "Short- and Intermediate-term Carcinogenicity Testing – a Review. Part 1: The Prototypes Mouse Skin Tumour Assay and Rat Liver Focus Assay", *Food Chem Toxicol,* 36: 979–995.

Enzmann, H., M. Iatropoulos, K.D. Brunnemann, E. Bomhard, H.J. Ahr, G. Schlueter and G.M. Williams (1998b), "Short- and Intermediate-term Carcinogenicity Testing – A Review. Part 2: Available Experimental Models", *Food Chem Toxicol,* 36: 997–1013.

ESAC [ECVAM Scientific Advisory Committee] (2006), *Statement on Dog Toxicity Studies*. http://ecvam.jrc.it/publication/ESAC25_statement_DOG_20061207.pdf

Gariboldi, M., G. Manenti, F. Canzian, F.S. Falvella, M.A. Pierotti, G. Della Porta, G. Binelli and T.A. Dragani (1993), "Chromosome Mapping of Murine Susceptibility Loci to Liver Carcinogenesis", *Cancer Res.*, 53, 209–211.

Giknis, M.L.A. and C.B. Clifford (2010), *Spontaneous Neoplastic Lesions in the Crl:CD-1 (ICR) Mouse in Control Groups from 18-Month to 2 Year Studies*, Charles River Laboratories, Wilmington, MA.

Griffiths, S.A., C. Parkinson, J.A.N. McAuslane and C.E. Lumley (1994), "The Utility of the Second Rodent Species in the Carcinogenicity Testing of Pharmaceuticals", *The Toxicologist,* 14(1):214.

Gold, L.S., T.H. Slone (1993), "Prediction of Carcinogenicity from Two Versus Four Sex-species Groups in the Carcinogenic Potency Database", *Journal of Toxicology and Environmental Health*, 39:143-57.

Goldsworthy, T.L., M.H. Hanigan and H.C. Pitot (1986), "Models of Hepatocarcinogenesis in the Rat-contrasts and Comparisons", *Crit Rev Toxicol*, 17: 61–89.

Huff, J. and J. Haseman (1991), "Long-term Chemical Carcinogenesis Experiments for Identifying Potential Human Cancer Hazards: Collective Database of the National Cancer Institute and the National Toxicology Program (1976-1991)", *Environmental Health Perspectives*, 96:23-31.

IARC (1992), *Mechanisms of Carcinogenesis in Risk Identification*, Vainio, H., P. Magee, D. McGregor and A.J. McMichael Eds, IARC Scientific Publications 116, Lyon.

IARC (1999), *The Use of Short- and Medium-term Tests for Carcinogens & Data on Genetic Effects in Carcinogenic Hazard Evaluation*, McGregor, D.B., J.M. Rice and S. Vennitt Eds, IARC Scientific Publications 146, Lyon.

ICH (International Conference on Harmonisation) (1997), *ICH Guideline - S1B Testing for Carcinogenicity of Pharmaceuticals*.

ICH (International Conference on Harmonisation) (2009), *ICH Guideline - M3R2 Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals*.

Ito, N., K. Imaida, R. Hasegawa and H. Tsuda (1989), "Rapid Bioassay Methods for Carcinogens and Modifiers of Hepatocarcinogenesis", *Crit Rev Toxicol*, 19: 385–415.

Keenan, K.P. (1996), "Commentary: the Uncontrolled Variable in Risk Assessment: ad Libitum Overfed Rodents-Fat, Facts and Fiction", *Toxicol Pathol*, 24(3): 376 – 383.

King-Herbert, A.P., R.C. Sills and J.R. Bucher (2010), "Commentary: Update on Animal Models for NTP Studies", *Toxicol Pathol*, 38, 180-181.

Kobel, W., I. Fegert, R.L. Billington, K. Bentley, W. Bomann, P. Botham, B. Stahl, B. van Ravenzwaay and H. Spielmann (2010), "A 1-year Toxicity Study in Dogs is no Longer a Scientifically Justifiable Core Data Requirement for the Safety Assessment of Pesticides", *Critical Reviews in Toxicology*, 2010; 40(1): 1–15.

Manenti, G., G. Binelli, M. Gariboldi, F. Canzian, L. De Gregorio, F.S. Falvella, T.A. Dragani and M.A. Pierotti (1994), "Multiple Loci Affect Genetic Predisposition to Hepatocarcinogenesis in Mice", *Genomics*, 23, 118–124.

Manenti, G., F. Galbiati, S. Noci and T.A. Dragani (2003), "Outbred CD-1 Mice Carry the Susceptibility Allele at the Pulmonary Adenoma Susceptibility 1 (Pas1) Locus" *Carcinogenesis*, 24(6):1143-8.

Maronpot, R.R., M.B. Shimkin, H.P. Witschi, L.H. Smith and J.M. Cline (1986), "Strain A Mouse Pulmonary Tumor Test Results for Chemicals Previously Tested in the National Cancer Institute Carcinogenicity Tests", *J Natl Cancer Inst,* 76: 1101–1112.

Marsman, D.S., and J.A. Popp (1994), "Biological Potential of Basophilic Hepatocellular Foci and Hepatic Adenoma Induced by the Peroxisome Proliferator, Wy-14,643" *Carcinogenesis,* 15: 111–117.

Meyer, O. (2003) "Testing and Assessment Strategies, Including Alternative and New Approaches", *Toxicology Letters,* 140-141:21-30.

NAS (2007), *Toxicity Testing in the 21st century: A vision and a Strategy, National Research Council of the National Academies*, The National Academies Press, Washington DC, ISBN 13: 978-0-309-10992.

Nohynek, G.J., L. Longeart, B. Geffray, J.P. Provost and A. Lodola (1993), "Fat, Frail and Dying Young: Survival, Body Weight and Pathology of the Charles River Sprague-Dawley-Derived Rat Prior to and Since the Introduction of the VAFR Variant in 1988", *Human and Experimental Toxicology*, 12(2): 87-98.

OECD (1998), *Repeated Dose 90-day Oral Toxicity Study in Non-Rodents*, OECD Guideline for the Testing of Chemicals No 409, OECD, Paris.

OECD (2000), *Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation,* Series on Testing and Assessment No. 19, OECD, Paris.

OECD (2009), *Detailed Review Paper on Transgenic Rodent Mutation Assays,* Series on Testing and Assessment, No. 103, OECD, Paris.

REACH (2006), *Guidance on Information Requirements and Chemical Safety Assessment, Part B: Hazard Assessment,* REACH Regulation (EC) No. 1907/2006. Available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_part_b_en.pdf?vers=20_10_08 .

Santos, J., P. Fernández-Navarro, M. Villa-Morales, L. González-Sánchez, J. Fernández-Piqueras (2008), "Genetically Modified Mouse Models in Cancer Studies", *Clinical and Translational Oncology*, Volume 10, issue 12, p. 794 - 803.

SCHER (2009), *The Need for Non-human Primates (NHP) in Biomedical Research, Production and Testing of Products and Devices*, Opinion of the Scientific Committee on Health and Environmental Risks.

Storer, R.D., F.D. Sistare, M.V. Reddy and J.J. DeGeorge (2010), "An Industry Perspective on the Utility of Short-Term Carcinogenicity Testing in Transgenic Mice in Pharmaceutical Development", *Toxicol Pathol,* 38: pp. 51-61.

Todd, G.C. (1986), "Induction and Reversibility of Thyroid Proliferative Changes in Rats Given an Antithyroid Compound", *Vet Pathol* 23: 110–117.

US EPA (1998c) Health Effects Test Guidelines OPPTS 870.4300 Combined Chronic Toxicity/Carcinogenicity. US Environmental Protection Agency, EPA 712-C-98-212, August 1998.

Usui, T., S.A. Griffiths and C.E.Lumley (1996), *The Utility of the Mouse for the Assessment of the Carcinogenic Potential of Pharmaceuticals*, In D'Arcy P.O.F. and D.W.G. Harron (eds), Proceedings of the Third International Conference on Harmonisation, Queen's University Press, Belfast. pp 279-284.

WHO (1990), *Principles for the Toxicological Assessment of Pesticide Residues in Food*, IPCS Environmental Health Criteria 104, World Health Organisation, Geneva.

Williams, G.M., M.F. Laspia and V.C. Dunkel (1982), "Reliability of the Hepatocyte Primary Culture/DNA Repair Test in Testing of Coded Carcinogens and Non-carcinogens", *Mutat Res,* 97: 359–370.

3.4 TOXICOKINETICS

167. Studies examining the toxicokinetics (TK) of a chemical substance are conducted to obtain adequate information on its absorption, distribution, biotransformation (i.e., metabolism) and excretion, to aid in relating concentration or dose to the observed toxicity, and to aid in understanding its mechanism of toxicity (OECD, 2010). Basic TK parameters determined from these studies will also provide information on the potential for accumulation of the test substance in tissues and/or organs and the potential for induction of biotransformation as a result of exposure to the test substance (OECD, 2010). Toxicokinetic studies may provide useful information for determining dose levels for toxicity studies (linear vs. non-linear kinetics), route of administration effects, bioavailability including differences in single versus repeat dose, internal dose, metabolism pathways (including generation of reactive intermediates) and the existence of saturation points of uptake, metabolism and/or excretion.

168. The specific objectives of a toxicokinetic study, as an adjunct to a chronic toxicity or carcinogenicity study, include the following (ICH, 1994):

- to describe the systemic exposure achieved in animals and its relationship to dose level and the time course of the toxicity study.
- to relate the exposure achieved in toxicity studies to toxicological findings and to contribute to the assessment of the relevance of these findings for other species i.e., humans/extrapolation.
- to provide information which, in conjunction with the toxicity findings, contributes to the design of subsequent toxicity studies including studies on MOA.

169. For the purpose of dose selection, TK studies are informative in indicating whether there is a “point of saturation" or saturation kinetics evident in the dose response curve (see also Section 3.1 on Dose selection).

170. The kinetics of absorption will determine the internal exposure achieved. The absorption and clearance of the compound and its metabolites will determine the systemic and target organ exposure resulting from a single dose and can be used to design the treatment regimen required to achieve a desired internal dose for either parent compound or major human metabolites. The effect of repeated exposures on absorption, metabolism, biotransformation, and clearance of a compound will provide information on the internal dose achieved during chronic exposure under conditions of the bioassay.

171. The bioavailability of test substance is often very dependent on the matrix it is administered in, e.g., due to the fat content. If this is the feed, there may be an interaction of the test substance with food matrix. The food composition may alter bioaccessibility. Therefore, exposure via food may provide different toxicokinetics compared to exposure via drinking water. In the few cases where administration in a chronic study is via gavage, it is important to realise that the composition of the gavage administration may influence bioavailability as well. These aspects should be considered when the results are used for risk assessment purposes.

172. As indicated in OECD TG 417 on Toxicokinetics (OECD, 2010), there are numerous studies that might be performed to evaluate the TK behaviour of a chemical for regulatory purposes. However, depending on particular regulatory needs or situations, not all of these possible studies may be necessary for the evaluation of a chemical. Flexibility, taking into consideration the characteristics of the substance being investigated, is needed in the design of toxicokinetic studies. In some cases, only a certain set of questions may need to be explored in order to address chemical-associated hazard and risk concerns. In some situations, TK data can be collected as part of the evaluation in other

toxicology studies. For other situations, additional and/or more extensive TK studies may be necessary, depending on regulatory needs and/or if new questions arise as part of chemical evaluation (see also Barton *et al.*, 2006).

173. In order to be of maximum utility in planning the design of a chronic toxicity or carcinogenicity study, particularly in the selection of dose levels, TK studies should be carried out, or data should be available, in the same species used in the long-term study and should preferably be performed using the same route and, where appropriate, the same vehicle as that used in the other toxicity studies. It should be noted however that such data may not be readily available for all chemicals, as they are not required under all regulatory schemes.

174. While single dose TK studies may provide useful information on absorption, distribution, metabolism and excretion of the test substance, the information most relevant in the planning and the execution of a chronic toxicity or carcinogenicity study will come from a repeat-dose toxicokinetic study over an extended period. As noted in OECD TG 417, repeated administration of the test substance may be needed to address more fully the potential for accumulation and/or persistence or changes in TK, or as required by a competent authority.

175. In addition to data from dedicated toxicokinetic studies such as OECD TG 417, useful information on repeat-dose toxicokinetics may be generated as part of a chronic toxicity (TG 452) or carcinogenicity (TG 451) study, or the combined chronic toxicity/carcinogenicity study (TG 453). In many cases, it is not necessary to add additional satellite animals into the study design for the purpose of providing excreta and blood samples for toxicokinetic analysis, since a minimum required number of blood samples to calculate representative toxicokinetics can be obtained from the study animals without compromising the outcome of the toxicity study (Saghir *et al.*, 2006). In the case of investigations carried out using the main study animals, the volume and number of blood samples which can be obtained per animal may, however, be limited by the stress imposed on the animals and the potential effects of repeated sampling on animal health and/or physiology. For more extensive TK investigations satellite groups or a separate TK study will be necessary. In the case of investigations using satellite groups, the quantity of test substance and in some cases metabolites excreted in urine, feces, and expired air should be measured on at least two time points on day 1 of collection (one of which should be at 24 hours post-dose), one at 48 hours, one at 7 days, one at 3 months, one at 12 months and one at termination. Blood samples should be taken from the satellite animals (and also in the case of an independent TK study) at suitable time points. Comparison of the area-under-curve (AUC) on Day 1 and the last day is used to assess issues such as accumulation and induction or inhibition of biotransformation, affecting AUC.

176. Guidance on toxicokinetic investigations following administration of test substance by the oral route(s) is given in the OECD TG 417. Information how to assess absorption following administration of test substance by the dermal route is given in the OECD TG 427 (*in vitro*) and TG428 (*in vivo*).

177. With respect to plasma levels of the test chemical measured in toxicity studies, an important point to note is that in rats there is a marked influence of sex hormones on liver biotransformation processes (see e.g., Chhabra & Fouts, 1974). In general, male rats metabolise xenobiotics (as well as endogenous substrates) faster than females, a finding not generally seen in other species. Thus rat studies may exhibit sex differences in plasma kinetics and in clinical and toxicological effects of the test chemical. Due to these differences in CYP3A expression pattern between males and females in rat and other species, these findings may not be relevant to human exposure.

178. As indicated in OECD TG 417 on Toxicokinetics (OECD, 2010), all available information on the test substance and relevant metabolites and analogs should be considered by the testing laboratory prior to conducting an additional toxicokinetic study in order to enhance study quality and

minimise animal usage. This could include data from other relevant test methods (*in vivo* studies, *in vitro* studies, and/or *in silico* evaluations). Physicochemical properties, such as octanol-water partition coefficient (expressed as log P_{OW}), pKa, water solubility, vapour pressure, and molecular weight of a chemical may be useful for study planning and interpretation of results. They can be determined using appropriate methods as described in the relevant OECD Test Guidelines.

179. The revised TG 417 also provides guidance on use of supplemental approaches in addition to the *in vivo* studies described in the preceding paragraphs that can provide useful information on absorption, distribution, metabolism and excretion (OECD, 2010). For example, use of freshly isolated or cultured hepatocytes and subcellular fractions (e.g., microsomes and cytosol or S9 fraction) from liver can provide useful information on metabolism of the test substance. Local metabolism in the target organ, e.g., lung, may be of interest for risk assessment. For these purposes, microsomal fractions of target tissues may be useful. Studies with microsomes may be useful to address potential gender and life-stage differences and characterize metabolic rates (Km and Vmax) which can aid in the assessment of dose dependency of metabolism in relation to exposure levels. In addition microsomes may be useful to identify the specific microsomal enzymes involved in the metabolism of the substance which can be relevant in species extrapolation.

180. In certain circumstances and under appropriate conditions, subcellular fractions coming from human tissues might be considered for use in determining potential species differences in biotransformation. Primary cell cultures from liver cells and fresh tissue slices may be used to address similar questions as with liver microsomes. In certain cases, it may be possible to answer specific questions using cell lines with defined expression of the relevant enzyme or engineered cell lines. It may also be useful to study the inhibition and induction of specific cytochrome P450 isozymes (e.g., CYP1A2, 2A1, and others) and/or phase II enzymes by the parent compound using *in vitro* studies. Information obtained may have utility for similarly structured compounds (OECD, 2010). The potential for induction of biotransformation can be examined by using liver subcellular fractions (e.g., microsomes and cytosol) of animals pretreated with the substance of interest, *in vitro* via hepatocyte induction studies or from specific cell lines expressing relevant enzymes (OECD, 2010).

181. The results from *in vitro* investigations may also have utility in the development of PBTK (physiologically-based toxicokinetic) models (Loizou *et al*., 2008). *In vitro* dermal absorption studies may provide supplemental information to characterize absorption (OECD, 2004b, 2004c).

182. Toxicokinetic models such as PBTK modelling may have utility for various aspects of hazard and risk assessment as for example in the prediction of systemic exposure and internal tissue dose. A PBTK model comprises an independent structural mathematical model, comprising the tissues and organs of the body with each perfused by, and connected via, the blood circulatory system. PBTK modelling may be used to predict the target tissue dose of the parent chemical or its reactive metabolite. Information derived from PBTK modelling experiments may aid in the comparison of biotransformation and toxicokinetics'of a test substance and/or its metabolites and may provide a basis for extrapolation across species or dosing patterns. Such experiments may also provide estimates of relevant internal tissue dose which might be important to the hazard or risk assessment process (Andersen, 2003; US EPA 2006; Nielsen *et al*., 2008, Clewell and Clewell, 2008). Furthermore, specific questions on mode of action (see Chapter 2) may be addressed.

183. Data useful for developing PBTK models for a chemical in any given species include 1) partition coefficients, 2) metabolic rate, 3) route-specific absorption parameters and 4) *in vivo* kinetic data for model evaluation (e.g., clearance parameters for relevant (> 10 %) excretion pathways, Km and Vmax for metabolism) (OECD, 2010). The experimental data used in model development should be generated with scientifically sound methods. The model predictions should be evaluated using experimental data as much as possible (IPCS, 2010). Chemical- and species-specific parameters such

as absorption rates, blood-tissue partitioning and metabolic rate constants are often determined to facilitate development of physiologically-based models (IPCS, 2010).

184. The ICH note for guidance on the Assessment of Systemic Exposure in Toxicity Studies provides additional guidance on the value of TK data in dose selection in carcinogenicity studies (ICH, 1994). The ICH note for guidance emphasises the need to estimate systemic exposure to the parent compound and/or metabolite(s) at appropriate dose levels via TK studies and at various stages of a carcinogenicity study, in order to ensure that the findings of the study can be interpreted in relation to the comparative exposure for the animal model and humans. The note for guidance notes that increases in exposure may arise unexpectedly as a result of non-linear kinetics due to saturation of a clearance or absorption process. Increasing systemic/internal exposure may also occur during the course of a study for those compounds which have a particularly long plasma half-life. With particular reference to administration of the test compound by oral gavage careful attention should also be paid to compounds which achieve high Cmax values over comparatively short time periods within the dosing interval. Conversely, unexpectedly low internal doses may occur during a study as a result of enzyme induction over time.

REFERENCES

Andersen, M.E. (2003), "Toxicokinetic Modelling and its Application in Chemical Risk Assessment", *Toxicol Lett*, 138, 9-27.

Barton, H.A., T.P. Pastoor, K. Baetcke, J.E. Chambers, J. Diliberto, N.G. Doerrer, J.H. Driver, C.E. Hastings, S. Iyengar, R. Krieger, B. Stahl, C. Timchalk (2006), "The Acquisition and Application of Absorption, Distribution, Metabolism, and Excretion (ADME) Data in Agricultural Chemical Safety Assessments", *Crit Rev Toxicol,* Vol. 36: 9-35.

Chhabra, R.S. and J.R. Fouts (1974), "Sex Differences in the Metabolism of Xenobiotics by Extra-Hepatic Tissues in Rats", *Drug Metab Dispos*, 6: 375-379.

Clewell, R.A. and H.J. Clewell III (2008), "Development and Specification of Physiologically Based Pharmacokinetic Models for Use in Risk Assessment", *Regul Toxicol Pharmacol* 50, 129-143.

ICH (International Conference on Harmonisation) (1994), *ICH Harmonised Tripartite Guideline: Note for Guidance on Toxicokinetics: Assessment of Systemic Exposure in Toxicity Studies. S3A* Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S3A/Step4/S3A_Guideline.pdf

IPCS (2010), *Characterizing and Applying Physiologically-based Pharmacokinetic Models in Risk Assessment*, World Health Organization, International Programme on Chemical Safety, Geneva. http://www.who.int/ipcs/methods/harmonization/areas/pbpk_models.pdf

Loizou, G., M. Spendiff, H.A. Barton, J. Bessems, F.Y. Bois, M. Bouvier d'Yvoire, H. Buist, H.J. Clewell 3rd, B. Meek, U. Gundert-Remy, G Goerlitz and W. Schmitt (2008), "Development of Good Modelling Practice for Physiologically Based Pharmacokinetic Models for Use in Risk Assessment: the First Steps", *Regulatory Toxicology and Pharmacology,* 50, 400 – 411.

Nielsen, E., G. Østergaard and J.C. Larsen (2008), *Toxicological Risk Assessment of Chemicals*, Informa Healthcare, US, ISBN-13: 978-0-8493-7265-0.

OECD (2004a), *Skin absorption: in vivo method,* OECD Guideline for the Testing of Chemicals No 427, OECD, Paris.

OECD (2004b), *Skin Absorption: In Vitro Method,* OECD Guideline for Testing of Chemicals No 428, OECD, Paris.

OECD (2004c), *Guidance Document for the conduct of skin absorption studies,* Series on Testing and Assessment, OECD, Paris.

OECD (2010), *Toxicokinetics*, OECD Guideline for the Testing of Chemicals No 417, OECD, Paris.

Saghir, S.A., A.L. Mendrala, M.J. Bartels, S.J. Day, S.C. Hansen, J.M. Sushynski and J.S. Bus (2006), "Strategies to Assess Systemic Exposure of Chemicals in Subchronic/Chronic Diet and Drinking Water Studies", *Toxicology and Applied Pharmacology*, 211, 245 – 260.

US EPA (2006), *Approaches for the Application of Physiologically-Based Pharmacokinetic (PBPK) Models and Supporting Data in Risk Assessment*.
Available at http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=157668 .

3.5 HOUSING, FEEDING, HANDLING OF ANIMALS AND EXPERIMENTAL PROCEDURES

185. Many national and international standards have been developed for animal care including housing, feeding health and handling, e.g., NRC (1995, 1996), Council of Europe (2006), the European Community (EEC, 1986) the Society for Laboratory Animal Science (GV-SOLAS, 1988), the Victorian Government Department of Primary Industries (2004). The general principles outlined in these guidelines are similar, and in conducting a chronic toxicity or carcinogenicity study, those guidelines applicable at a national level should be followed.

186. An overarching principle is that the particular needs of given species and strains must take precedence and that adherence to guidelines should never replace close observation of the particular animals involved, continued throughout their lives (Council of Europe, 1997). Provision of exhaustive guidelines for all species and strains is difficult to achieve and local initiatives for improving housing conditions should be taken whenever possible. Appendix A of the Council of Europe Convention "Guidelines for accommodation and care of animals", does however provide detailed guidance on these issues, including aspects such as design and maintenance of the test facilities (Appendix A, Council of Europe, 2006). It should be consulted for in-depth information.

3.5.1 Housing

187. Taken as an example, the Council of Europe Convention (Appendix A, 2006) states on housing that special relevance should be given to the enrichment of the environment of the respective species according to their needs for social interaction, activity-related use of the space, appropriate stimuli and materials. In a review of laboratory environments and rodents' behavioural needs, Balcombe (2006) notes that there is growing recognition of the inherent problems of depriving rodents the space and resources to carry out natural behaviours, such as exploring, foraging, running, escaping hiding and hygiene maintenance. The author reports a recent survey of animal facilities at the US National Institutes of Health which indicates that a slight majority of rats and mice at these facilities are now being provided with nesting and structural (shelter) enrichment (Hutchinson *et al*., 2005). Other indicators that rodent housing conditions are improving include the availability of commercially produced resources for nesting, shelter, gnawing and play (Key, 2004), and a sharp rise since the late 1980s in the number of citations using keywords 'environmental enrichment' and 'rodent' (Hutchinson *et al*., 2005). Considering that two decades ago environmental rodent enrichment was scarcely being discussed, the author notes that these are laudable trends (Balcombe, 2006).

188. The Council of Europe Convention Appendix A (2006) recommendations on housing for rodents are as follows:

- Rodent species other than guinea pigs should be kept in cages made of easy to clean material and their design should allow proper inspection of the animals without unnecessarily disturbing them;

- The cages should be provided with solid floors with bedding instead of grid floors, unless there is good reason to have alternatives;

- Gregarious species should normally be group-housed, although it may be difficult to achieve stable and harmonious groups of male mice, and also female hamsters;

- Where the experimental procedures or welfare requirements make group-housing impossible, consideration should be given to accommodating animals of the same species within sight, sound or smell of one another;

- Encouragement should be given to break up the interior space of a cage by introducing objects such as platforms, tubes, boxes, etc. and attempts should be made to provide environmental enrichment with objects to explore, carry or transform, unless negative effects are observed on welfare or on the intended scientific use;

- High hygiene standards should be maintained. However, it may be advisable to maintain odour patterns left by the animals;

- Special attention should be paid to ensuring that the lighting intensity particularly on the top row of cages is not too high. Maximum light intensity should not exceed 350 Lux measured 1 metre from the floor. Provision should be made for shaded areas within the cage to allow the animals to withdraw.

189. The Convention makes specific recommendations for size of caging and stocking densities, dependent on the size/weight of the animals. In relation to environmental conditions, the Convention provides specific recommendations for temperature, humidity and ventilation for each species of laboratory animal covered in the guidelines. Those for rodents are in line with those indicated in the OECD Test Guidelines, as outlined in the next paragraph.

190. The Test Guidelines make some specific recommendations for housing of rodents only, including the recommendation (in line with that of the Convention) that animals may be housed individually e.g., when the test substance is administered via the dermal route, in order to prevent grooming behaviours and oral ingestion, or be caged in small groups of the same sex; individual housing should be considered only if scientifically justified. Animals may be group-caged by sex, but the number of animals per cage must not interfere with clear observation of each animal. The biological properties of the test substance or toxic effects may indicate a need for individual caging. Rodents should be housed individually in dermal studies and during exposure in inhalation studies.

191. The Test Guidelines also specify that cages should be arranged in such a way that possible effects due to cage placement are minimised. For rodents, the temperature in the experimental animal room should be 22°C (± 3°C). Although the relative humidity should be at least 30% and preferably not exceed 70% other than during room cleaning, the aim should be 50-60%. Lighting should be artificial, the sequence being 12 hours light, 12 hours dark. These recommendations, together with those of the Convention should be applied in any rodent chronic toxicity or carcinogenicity study conducted according to the Test Guidelines.

192. As indicated in Section 3.3, although rodents (rats or mice) are the main species used in chronic toxicity or carcinogenicity studies, other species, dog in particular, may be used on some occasions. As for rodents, specific guidelines for the care of dogs including housing, feeding health and handling have been developed, e.g., Council of Europe (2006). In relation to the housing of dogs, the Council of Europe Convention recommends that:

- Dogs should be housed in socially harmonious groups, unless the experimental procedures or welfare requirements make this impossible;

- Dogs should be exercised at least daily. Under no circumstances should dogs be caged without exercise for more than 14 days. Preferably, dogs should be exercised with other dogs.

- Dog pens should allow some privacy for the animals. They should include playthings and structures, including elevated platforms.
- Solid floors should be used for dogs. The materials, design and construction of slatted or perforated floors should provide surfaces which do not produce welfare problems such as irritation or injury of the feet or toes, blistering, etc. (these must be prevented at all times), and should supply a solid resting area.
- Temperature in dog studies should be held within a range of 15-21°C, light period between 10 and 12 hours a day, and humidity 40-70%.

193. In relation to the housing of minipigs, pigs should be housed in socially harmonious groups, unless the experimental procedures or welfare requirements make this impossible (Ellegaard *et al*., 2010).

- Solid floors should be used for pigs. To satisfy rooting behaviour of pigs the use of bedding is recommended. The use of bedding serves as a nutritional substrate, as well as providing environmental enrichment, especially in experimental units where the possibility to move around is limited. Where slatted or perforated floors are used, the materials, design and construction should provide surfaces which do not produce welfare problems such as irritation or injury of the feet or toes, blistering, etc. (these must be prevented at all times), and should supply a solid resting area.
- Pigs are highly sensitive to environmental temperature. The recommended temperature in pig studies should be held within a range of 17.5 to 24°C when 14-16 weeks old and 34-36 weeks of age respectively (dependent on use of bedding; higher temperature for piglets), light period at a minimum of 8 hours a day, and humidity 45-75%.

3.5.2 Feeding

194. In both humans and laboratory animals, diet has a direct bearing on health, and many neoplastic and non-neoplastic diseases are caused (or prevented) by dietary factors, including variations in the composition and amount of feed consumed. The association in rats of caloric consumption, the spontaneous formation of tumours and life span is well established. Although the zero-dose group may be expected to control for the influence of diet, dietary constituents may still profoundly affect the outcome of an experiment (OECD, 2002).

195. A nutritionally-balanced diet is important both for the welfare of laboratory animals and to ensure that experimental results are not biased by unintentional nutritional factors (NRC, 1995). The US National Research Council provides detailed guidance on the nutritional requirements of a wide range of laboratory animals, with detailed information on essential nutrients and other considerations for each species (NRC, 1995). The NRC guidance emphasizes that feed palatability and intake, nutrient absorption and utilization, and excretion can be affected by physicochemical characteristics of feeds such as physical form, sensory properties, naturally-occurring refractory or anti-nutritive substances, chemical contaminants, and conditions of storage (NRC, 1995). Many biological factors also affect nutritional requirements, including genetic differences between species and strains, stage of life of the animals, environmental influences (e.g., diurnal rhythms, temperature etc.), housing and microbiological status (NRC, 1995). Detailed information is also given on diet formulation for natural-ingredient diets, purified and chemically-defined diets, and on manufacture and storage procedures and other considerations (NRC, 1995).

196. The Test Guidelines state that rodents should be fed and watered ad libitum with food replaced at least weekly. Conventional laboratory diets should normally be used/ are normally used. The diet should meet all the nutritional requirements of the species tested and the content of dietary contaminants, including but not limited to pesticide residues, persistent organic pollutants, phytoestrogens, heavy metals and mycotoxins, that might influence the outcome of the test, should be as low as possible. Control and test animals should be fed from the same batch and lot. Analytical information on the nutrient and dietary contaminant levels should be generated periodically, at the beginning of the study and whenever there is a change in the batch used, and should be included in the final report. Analytical information on the drinking water used in the study should similarly be provided. The choice of diet may be influenced by the need to ensure a suitable admixture of a test substance in the diet and to meet the nutritional requirements of the animals when the test substance is administered by the dietary route.

197. As noted in section 3.2.1, the concentration of the test substance in the feed should not normally exceed an upper limit of 5% of the total diet (FDA, 1982, Borzelleca, 1992), although higher levels are feasible (e.g., when testing carbohydrates or proteins) as long as the diet is adapted to be nutritionally adequately, e.g., incorporated, at the expense of other components, in a purified diet. Section 3.2.1 also discusses the problems associated with the palatability of diet (or drinking water) containing test substances affecting the taste and/or smell of the food. If this is marked, it may be necessary to introduce into the study design an additional control group, pair fed (i.e., having matched food intake) in parallel with the high dietary level test group.

198. An important aspect of the feeding regime used in chronic toxicity and carcinogenicity is the recognized effect on study outcome of feeding *ad libitum*. Traditionally, maximal growth and reproduction have been used as criteria for the evaluation of laboratory animal diets (NRC, 1995). However, evidence from a number of studies indicates that restricting the caloric intake of laboratory animals may have beneficial effects on life span, the incidence and severity of degenerative diseases, and the onset and incidence of neoplasia (Weindruch and Walford, 1988; Yu, 1994; Keenan *et al*., 1997). Based on these results, allowing animals to eat *ad libitum* to produce maximum growth and reproduction may not be consistent with objectives of long-term toxicological and aging studies (NRC, 1995). Overfeeding by *ad libitum* food consumption is generally considered to be the most significant, uncontrolled variable affecting the outcome of the current rodent bioassay, and in particular, the correlation of food consumption, the resultant adult body weight and the 2-year survival in Sprague-Dawley rats is highly significant (Keenan *et al*., 1997). However, it will probably take years to introduce dietary restriction into national and international test guidelines for toxicity testing because of concern that the delayed occurrence of, for example, cancer reflects a decrease in the sensitivity of the carcinogenicity test in detecting the carcinogenic potential of tested chemical and because the considerable database on historical control is based on data from *ad libitum* feeding studies (Meyer *et al.*, 2003). Species and strain differences in survival are discussed further in Section 3.3.

199. At a practical (experimental) level however, restriction of the caloric intake of laboratory animals is not straightforward. It may disrupt normal diurnal eating rhythms and is not compatible with group housing. It is also important to achieve caloric restriction of test animals without producing unintended nutrient deficiencies (NRC, 1995). Elevation of nutrient concentrations in the diet may be necessary to ensure that the nutrient intake of animals whose eating is restricted is comparable to that of animals allowed to eat *ad libitum.* There is, however, relatively little information available about the extent to which caloric restriction affects nutrient requirements (NRC, 1995). Since rats regulate their food intake according to caloric intake, the mineral and vitamin etc. content of the diet should be adjusted to "caloric density".

200. As already noted, the US National Research Council provides detailed guidance on the nutritional requirements of a wide range of species other than laboratory rodents, including dogs and

rabbits (NRC, 1995). In the case of a chronic toxicity or carcinogenicity study involving animals other than rodents, this guidance should be consulted for information regarding feeding.

3.5.3 Handling, Health Surveillance and Experimental Procedures

201. The quality of care provided in the laboratory may influence not only growth rate and welfare, but also the quality and outcome of experimental procedures (Council of Europe Convention, 2006). The animals should be accustomed to competent and confident handling during routine husbandry and procedures; this will reduce stress both to animals and personnel. Hurst and West (2010) note that *"consistent use of handling methods that do not induce strong anxiety responses will minimise confounding responses due to routine handling before and during experiments, reducing the need to standardise handling experience and timing"*. These authors demonstrated that picking up mice by the tail induced high anxiety and aversion, while use of tunnels or open hand led to voluntary approach, low anxiety and acceptance of physical restraint. Non-rodent species such as dogs, animals should be handled or be in social contact with humans on a regular basis. The behaviour of an animal during handling and the performance of experimental procedures depend to a considerable extent on the confidence and competence of its handler. Good technique should be unhurried, sympathetic and gentle but firm and safe for the animal and operator. All personnel should be appropriately educated and trained, and records of training maintained.

202. A strategy should be in place in all establishments to ensure that an appropriate health status is maintained, which safeguards animal welfare and meets scientific requirements (Council of Europe Convention, 2006). This strategy should include a microbiological surveillance programme, plans for dealing with health breakdowns, and should define health parameters and procedures for the introduction of new animals, e.g., quarantining. Supervision of the accommodation and care by a veterinarian or other competent person is essential.

203. In relation to the experimental phase of a chronic toxicity or carcinogenicity study, as indicated in the Test Guidelines, the animals selected for the study should have been acclimated to laboratory conditions for at least 7 days and should not have been subjected to previous experimental procedures. A period of acclimatisation is needed to allow animals to recover from transport stress, to become accustomed to a new environment and to husbandry and care practices, and to ensure that their health status is sound. The test animals should be characterized as to species, strain, source, sex, weight and age. Each animal should be assigned a unique identification number, and permanently marked with this number by tattooing, microchip implant, or other suitable method. The method chosen should be reliable and cause the minimum pain and discomfort to the animal when applied and in the long-term. Staff should be trained in carrying out the identification and marking techniques, and sedatives or local anaesthetics and analgesics should be used if necessary.

204. At the commencement of the study, the weight variation for each sex of animal used should be minimal and not exceed ± 20 % of the mean weight of all the animals within the study, separately for each sex. Animals should be randomly assigned to the control and treatment groups. After randomisation, there should be no significant differences in mean body weights between groups within each sex. If there are statistically significant differences, then the randomisation step should be repeated, if possible.

205. The animals should be inspected regularly throughout the study, at least daily by a trained person, to ensure that all sick or injured animals are identified and appropriate action taken. Regular health monitoring should be carried out. The Test Guidelines specify that all animals should be checked for morbidity or mortality, usually at the beginning and the end of each day. Animals should additionally be checked once a day following dosing in the case of gavage studies, for specific signs

of toxicological relevance, taking into consideration the peak period of anticipated effects after dosing in the case of gavage administration. Particular attention should be paid to tumour development. The time of tumour onset, location, dimensions, appearance, and progression of each grossly visible or palpable tumour should be recorded. Body weights and food/water consumption and food efficiency should be assessed and recorded at the intervals specified in the guidelines.

206. At the end of the study, for interim kills and in the case of animals found sick or moribund during the study, the animals should be humanely killed. For non-scheduled killing i.e., for animals showing clinical sign of pain, suffering or distress, OECD Guidance Document 19 on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoint for Experimental Animals Used in Safety Evaluations should be followed (OECD, 2000). All humane methods of killing animals require expertise, which can only be attained by appropriate training. Animals should be killed using a method that adheres to the principles set by the European Commission Recommendations for the euthanasia of experimental animals (Part 1 and Part 2) (EEC, 1986). A deeply unconscious animal can be exsanguinated, but drugs which paralyse muscles before unconsciousness occurs, drugs with curariform effects and electrocution without passage of current through the brain, should not be used without prior anaesthesia. Disposal should not be allowed until death has been confirmed.

207. Records of source, use and final disposal of all animals bred, kept for breeding, or for subsequent supply for use in scientific procedures should be used not only for statistical purposes but, in conjunction with health and breeding records, as indicators of animal welfare and for husbandry and planning purposes.

REFERENCES

Balcome, J.P. (2006), "Laboratory Environments and Rodents' Behavioural Needs, a Review", *Laboratory Animals*, 40, 217–235.

Borzelleca, J. F. (1992), "Macronutrient Substitutes: Safety Evaluation", *Regulatory Toxicology and Pharmacology*, 16, 253-264.

Council of Europe (1986, adopted 2005), *European Convention for the Protection of Vertebrate Animals used for Experimental and other Scientific Purposes*, CETS No.123. Appendix A of the Convention: Guidelines for accommodation and care of animals (Article 5 of the Convention) approved by the Multilateral Consultation, 15th June, 2006. Available at: http://conventions.coe.int/Treaty/EN/Treaties/PDF/123-Arev.pdf

EEC (1986), *Council Directive 86/609/EEC on the approximation of laws, regulations, and administrative provisions of the Member States regarding the protection of animals used for experimental and other scientific purposes*, Official Journal 29, L358, 18 December 1986.

Ellegaard, L., A. Cunningham, S. Edwards, N. Grand, T. Nevalainen, M. Prescott and T. Schuurman (2010), "Welfare of the Minipig with Special Reference to Use in Regulatory Toxicology Studies, *Journal of Pharmacological and Toxicological Methods*, 62, 167-183.

FDA (1982), *Toxicological principles for the safety assessment of direct food additives and color additives used in food*, US Food and Drug Administration, Washington, DC.

GV-SOLAS (Society for Laboratory Animal Science, Gesellschaft für Versuchstierkunde) (1988), Publication *on the Planning and Structure of Animal Facilities for Institutes Performing Animal Experiments*, ISBN 3-906255-04-2. http://www.gv-solas.de/publ/heft1_1988.pdf

Hurst, J.L. and R.S. West (2010), "Taming Anxiety in Laboratory Mice", *Nature Methods*, 7(10), 825-826.

Hutchinson, E., A. Avery and S. VandeWoude (2005), "Environmental Enrichment for Laboratory Rodents", *ILAR Journal*, 46, 148–61.

Keenan, K.P., G.C. Ballam, R. Dixit**,** K.A. Soper, P. Laroque, B.A. Mattson, S.P. Adams, J.B. Coleman (1997), "The Effects of Diet, Overfeeding and Moderate Dietary Restriction on Sprague-Dawley Rat Survival, Disease and Toxicology", *J Nutr,* 127, 851S-856S.

Key, D. (2004), "Environmental Enrichment Options for Laboratory Rats and Mice", *Lab Animal,* 33, 39–44.

Meyer, O. and O. Svendsen (2003), *Animal Models in Pharmacology and Toxicology*, Chapter 2 in Handbook of Laboratory Animal Science, second edition Volume II, Animal Models, J. Hau and G.L. Van Hoosier Jr, (eds), CRC Press.

NRC (National Research Council) (1995), *Nutrition Requirements of Laboratory Animals*, 4th revised edition, National Academy Press, Washington, D.C.

NRC (National Research Council) (1996), *Guide for Care and Use of Laboratory Animals*, National Academy Press, Washington, D.C.

OECD (2000), *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints,* Series on Testing and Assessment, No. 19, OECD, Paris.

OECD (2002), *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies,* Series on Testing and Assessment No. 35, OECD, Paris.

Victorian Government Department of Primary Industries, Australia (2004), *Code of Practice for the Housing and Care of Laboratory Mice, Rats, Guinea Pigs and Rabbits*. Available at: www.anu.edu.au/ro/ORI/Animal/code_of_practice_lab_animals.pdf

Weindruch, R. and R.L. Walford (1988), *The Retardation of Aging and Disease by Dietary Restriction*, Charles C Thomas, Springfield, IL.

Yu, B.P., ed. (1994), *Modulation of Aging Processes by Dietary Restriction*. CRC Press, Boca Raton, FL.

3.6 INVESTIGATIONS (INCLUDING HISTOPATHOLOGICAL GUIDANCE)

3.6.1 Introduction

208. This Section on investigations includes guidance on the design and conduct of pathological, histopathological and ophthalmoscopic investigations together with more general advice on the avoidance of bias during investigations. Sponsors and study directors should consider this guidance when planning and performing studies to reduce the likelihood of misleading or ambivalent findings that could result in inappropriate conclusions or the need to repeat studies and to optimize the use of animals for generation of data. It complements existing Guidance contained in OECD Guidance Document No. 35 on the analysis and evaluation of chronic toxicity and carcinogenicity studies. Guidance Document 35 should be consulted for detailed guidance on mortality, clinical observations, body weight changes, food and water consumption, absolute and relative organ weights, haematological, clinical and urinary measurements, post mortem observations and analysis of toxicokinetic and metabolism data. The Society of Toxicologic Pathology recommendations for organ weights have also been published (Sellers *et al.,* 2007).

209. Ophthalmoscopy is an extremely useful clinical technique that allows examination of the anterior and posterior of the eyeball (fundus), including the retina, optic disc, choroid, and blood vessels in the eye. Ophthalmoscopy can help to detect diseases of the eye, to diagnose other conditions or diseases that damage the eye, and can detect other diseases such as some brain tumours. In life ophthalmoscopy allows the progression of the disease to be followed in time and can indicate that additional sampling at necropsy would be informative.

210. Histopathology evaluation is an important part of the assessment of the adverse effects of chemicals on the whole organism. Conventional histological, histochemical and special staining techniques, together with electron microscopy can be used both to define the identity and morphology of tissue, cellular and subcellular structures and to indicate the chemical characteristics of their constituents. In addition to conventional special staining techniques and electron microscopy, new techniques involving immunohistochemistry, molecular biology and novel visualisation procedures can be applied to provide additional more objective methods. Information from such investigations can be used to obtain better functional and morphological characterization of induced alterations in tissues of the body, when needed.

3.6.2 Ophthalmoscopy

211. In toxicity studies, ophthalmoscopic evaluation should be conducted by a suitably trained and experienced individual, preferably using indirect fundoscopic examination *and* a slit-lamp evaluation. A topical mydryatic agent will normally be employed to facilitate ophthalmoscopic examination.

212. Animals should be examined pre-treatment and pre-terminally. Preferably, they will also be examined at least once during the course of the study. Dependent upon the nature of any abnormalities observed, additional ophthalmoscopic examination or other clinical investigations may be warranted. Ocular abnormalities may indicate additional tissue sampling at postmortem to enable histological examination of the ocular adnexa. When ocular abnormalities are detected by ophthalmoscopy and treatment-related causes cannot be confidently excluded, histopathological examination of the eye should attempt to identify the morphological correlate of the abnormality. If ocular abnormalities are focal, it may be appropriate to perform histological examination on additional sections. Appropriate sections from all groups should be taken to provide adequate controls.

213. The study report should contain an integrated interpretation of all treatment-related ocular findings (ophthalmoscopic, macroscopic and microscopic examinations), along with pertinent individual animal data.

214. It should be noted that rodent models are not suitable for the detection of disturbances of ocular pressure by tonometry or for monitoring for certain visual disturbances, such as dyschromatopsia.

3.6.3 Pathology and Histopathology

3.6.3.1 General Considerations

215. Pathology has an important role in toxicology since it provides information on the differences in tissue and organ morphology that establish the presence or absence of lesions and whether or not there are dose–effect relationships. Pathology data can facilitate the interpretation of other data, such as organ weight changes, clinical biochemistry or haematology findings (e.g., Krinke *et al.*, 1991), and evaluators should always make it clear whether there are any associations between pathological abnormalities and other findings of physiological significance. Nevertheless, not all changes in tissue morphology are accompanied by abnormalities in other parameters, and perturbation in organ biochemistry will not necessarily be accompanied by changes in the histological appearance of the affected organ(s). An overview of physiological and environmental factors that can complicate the interpretation of findings in a toxicity study may be found in the *Handbook of Toxicologic Pathology* (Bucci, 1991).

3.6.3.2 Sampling

216. The pathologist should ensure that standard sections of all appropriate tissues and organs are present on the slides to be evaluated. The use of standard sections helps to ensure comparable samples across all animals, thus reducing inconsistency. If flawed or incomplete specimens (eg. missing medulla of adrenal, pars distalis of pituitary, mucosa of intestine, the parathyroid), impair the pathologist's ability to detect, or evaluate treatment-related effects, it is the pathologist's responsibility to obtain recuts, to the extent possible, of those missing or inadequate tissues. Any irretrievable omissions should be taken into account in the interpretation of the data and discussed in the pathology narrative.

217. Procedures for tissue sampling and trimming to ensure optimal fixation, vary between laboratories. However, more standardised approaches to the selection of blocks, orientation of tissues and number of slides examined for rat and mouse organs and tissues in regulatory type toxicity studies are now available (Ruel-Fehlert *et al*., 2003; Kittel *et al*., 2004 and Morawietz *et al*., 2004) and are recommended. These publications are based on the experience of the European Registry of Industrial Toxicology Animal-Data (RITA) and North American Control Animal Database (NACAD). They are an extended revision of the trimming guidelines published by Bahnemann *et al*. (1995). The articles describe in detail the optimum localisation for tissue preparation, the sample size, the direction of sectioning and the number of sections to be prepared, organ by organ. However, existing information on the substances tested and gross findings at autopsy may indicate that samples from non standard locations should also be taken. If a test substance is administered by the dermal route or by inhalation route for example, samples should be taken from the site of application, in addition to the standard specimens.

3.6.3.3 Histopathology specimen processing and quality

218. High quality tissue specimens are necessary for histopathologic evaluation. For routine purposes, fixation in a suitable concentration of a formalin solution followed by processing and embedding in paraffin wax, and cutting of a suitable thickness of histological section is adequate for most tissues. Exceptions are the eyes and testes for which formalin is generally regarded as an inadequate fixative. In these instances, Bouin's or Davidson's fixatives are preferable, especially if evidence from shorter term studies indicated treatment-related testicular or ocular toxicity. Bouin's fluid is generally considered the best fixative for testis although there has been a move away from this because of the safety concerns of its picric acid content (Latendresse *et al*., 2002). Formalin, when used under stringent conditions is usually also a good fixative for many immunohistochemical and molecular biological techniques on tissue sections. Frozen sections are however usually needed for studies of enzyme activity or those requiring intact RNA. Tissues should be processed in a manner that reduces the potential for variation to be introduced among groups, as indicated in Section 3.6.5.

219. Special stains should be used where appropriate. Haematoxylin and eosin (H&E) remains the most widely used stain, supplemented, where appropriate by a Romanovsky stain for haemopoietic cells, Periodic-acid Schiff (PAS) stain for hepatic glycogen, glomerular basement membrane and the acrosome on testicular germ cells, trichrome and elastic stains for the myocardium, and blood vessels and oil red O applied to frozen sections for neutral lipids. While still not standard, immunohistochemical techniques are now widely used in special cases and the advantages and pitfalls of the techniques are discussed in detail by Greaves (2009). All specially stained tissues should be accompanied by a specimen known to be positive for the particular test to confirm the proper functioning of the technique. Where procedures other than the standard H & E staining are carried out, the specific fixation procedures appropriate to that method should be followed.

220. The use of larger semi-thin (1 to 3 µm thick) plastic or resin embedded sections is a technically demanding but cost effective compromise between electron microscopy and conventional light microscopy. It requires specialist equipment frequently not present in routine histology laboratories. Sometimes termed 'high resolution light microscopy', light microscopic evaluation of semi-thin sections provides a means of avoiding extensive use of the electron microscope because it can locate cytoplasmic organelles in a way sometimes not possible in a paraffin wax embedded material.

221. In view of the hazard to health of formalin, used for fixation, and xylene, used during processing, an increasing number of laboratories are adopting technologies which are free of both formalin and xylene. Nassiri *et al*. (2008) describes the utilization of a formalin-free fixation and processing system for tissue detection of two important biomarkers in breast cancer at the RNA and protein levels. Falkeholm *et al*. (2001) demonstrated that xylene-free histological sections are qualitatively on a par with conventional paraffin sections for routine diagnostic work. Overall, these techniques appear to have good safety profiles, provide excellent histology quality and are suitable for further analysis by immunohistochemistry and nucleic acid extraction. There can however be subtle differences between xylene-free and xylene sections thus the pathologist should be aware of the potential for these differences especially when reading between studies on the same compound where one study uses sections using xylene and the other is xylene-free.

222. In addition to histopathology and immunohistochemistry, tissue may be required for molecular genomic or transcriptomic analysis of biomarkers. This should be borne in mind when collecting and studying tissue. It is important therefore to be prepared to extract RNA, DNA and even protein from paraffin embedded material. This can be achieved using both formalin-containing and formalin-free fixed tissues. Hewitt *et al.* (2008) summarizes the current state-of-the-art of preanalytic

factors in tissue handling and processing as they impact the quality of RNA obtainable from formalin fixed paraffin-embedded tissue.

223. Artefacts may be produced at each of the following stages in the processing of tissue sections: before death, at postmortem or necropsy, during the fixation of tissues, during processing, paraffin embedding and microtomy, during the mounting of tissue sections onto glass slides, staining procedures and coverslipping. Some artefacts are easily distinguishable from normal or diseased tissue components but these render the task of the pathologist more difficult and great care should be taken to avoid their introduction. Some artefacts are difficult to distinguish from pathological changes and are thus of particular concern. McInnes (2005) sets out some of the more common artefacts that are most frequently encountered as a result of inadequacies in the preparation of microscopic tissue sections.

224. All slides evaluated should be identified with a unique code from which identifies the study number, animal number, slide (block) number and the tissue[s] present on the slide.

3.6.3.4 Histopathologic evaluation

225. Details of the nature of the test substance, results of any previous toxicity studies and known activities of this class of compounds should be made available to the pathologist before evaluation of the tissue slides begins. Knowledge of target organs and tissues and the types of changes previously encountered, even in different species, facilitates the evaluation of tissues and provides for the consistent use of terminology. Previous knowledge of target tissues should be utilized during the protocol development to determine whether special pathology procedures should be used in obtaining, fixing, processing, or staining of sections.

226. The study pathologist should have access to in-life clinical observations, organ and body weight data, haematology and clinical biochemistry data, and macroscopic findings of the postmortem examination for each animal in addition to complete information about the experimental design, characteristics of the animal (age, sex and strain) and husbandry of the study population. These include, but are not necessarily limited to: study protocol, including amendments and relevant deviations, species, strain, and age of animals, route, doses, and duration of dosing.

227. Metabolic, pharmacokinetic, or toxicokinetic information may be necessary for understanding patterns of change and interpreting differences in species responses.

228. In-life data (i.e., clinical signs, body weight changes, food consumption, ophthalmoscope findings etc.) from animals may help greatly in the identification of target organs and in understanding mechanisms of toxicity. Haematology, clinical chemistry, and urinalysis results also aid the identification of target organs and may contribute to an understanding of the mechanism of action. Results of special assays, such as hormone concentrations or enzyme induction, are equally important in locating morphologic changes and in the understanding of their significance.

229. Necropsy (gross) findings for individual animals must be available to the pathologist for lesion tracking and correlation with histopathology findings. The pathologist should also be aware of organ weight changes. Often histomorphologic correlates of altered weights can be identified.

230. All of the above-mentioned data, if available, should be provided to the pathologist at the time of the initial slide evaluation. Provision of information on previous findings for the individual animals may render the treatment status of the animal obvious. The pathologist may wish to examine samples from animals of the same treatment group for the same histopathological changes. This may

improve the efficiency of the process but comes at the price of inevitably introducing the potential for bias into the process. Careful consideration needs to be given to the balance between ensuring that subtle or rare changes are detected and the possible introduction of bias. This is discussed further in section 3.6.5.

231. Tissues may be evaluated animal by animal or organ by organ as the preference of the pathologist dictates. The animal by animal technique affords an encompassing overview of an animal's complete health status. The organ by organ technique allows more focused attention to changes and aids in the consistent grading of changes in a particular organ. Where concise terminology is inadequate to convey lesion complexity, detailed free text descriptions should be used to define the diagnostic term used for tabulation.

232. For common lesions in a species or strain of animal, it is important to know if the experimental treatment alters severity. The pathologist should use a severity grading system that allows for an appropriate severity classification, as treatment may affect the incidence (number of animals showing the pathology) or the severity of a particular pathology lesion. Toxicological lesions are frequently found in a continuous spectrum of severity. Therefore, severity grading systems should be: 1) definable, 2) reproducible, and 3) meaningful. A description of each of the various grades should be included in the narrative for target lesions where severity is critical to interpretation of the data. Photomicrographs may be helpful in conveying the severity differences for the grading system used. Well-defined severity grading systems greatly aid the pathology peer-review process. For carcinogenicity studies, it is the pathologist's responsibility to distinguish between hyperplasia, dysplasia, neoplasia and to classify tumours, where applicable, as 1) benign or malignant, and 2) primary or metastatic.

233. Computerised systems of recording findings help to organise the process of evaluation and ensure that all animals and all tissues have been examined, and that gross and microscopical correlation is followed. They also allow rapid, and reproducible, formation of the pathology tables needed to ensure appropriate interpretation of treatment-related effects.

3.6.3.5 Procedures to enhance the accuracy and consistency of histopathology

234. Histopathology is a descriptive and interpretive science and therefore includes an element of subjectivity. However, the pathologist should evaluate tissues as consistently as possible to avoid the introduction of artificial differences or between-group bias. Evaluation of all tissues within a study by one pathologist with consistent standards for detecting, naming, and grading tissue changes, facilitates the detection of differences induced by treatment. However, two or more pathologists are on rare occasions involved in evaluation of a study. In such circumstances, the utmost care must be taken to ensure that nomenclature and severity grading systems used by the contributing pathologists are harmonized to limit variation and that steps are taken to avoid bias (see section 3.6.5). For example, the situation where all of the samples in one dose group are evaluated by one pathologist and those in another dose group by another pathologist should be avoided. Standardized criteria and consistent terminology should be agreed upon for grading systems of common spontaneous and treatment-related findings. The use of a shared computer system that can define a study-specific lexicon for capturing data facilitates this process. Guidance on standard nomenclature is available in OECD Guidance Document No. 35.

235. Pathologists are aware of the phenomenon of "diagnostic drift." Drift refers to a gradual change in nomenclature or severity grading of lesions within a single study. Diagnostic drift usually develops from the increased awareness of a lesion by the pathologist and it is more of a problem in large studies with many animals and tissues requiring evaluation over a prolonged period of time. It is

a source of inconsistency that can negatively impact detection of treatment-related lesions or artefactually introduce apparent treatment-related effects where none exist and can erroneously affect the determination of no-effect-levels. When a pathologist becomes aware of drift in his/her selection of terminology or severity grading, he/she must re-evaluate the tissue(s) involved. Any suspected treatment-related effects should normally be re-evaluated through the use of a "blinding" or "masking" technique, where appropriate.

3.6.3.6 Image Capture

236. It is useful to capture images of key features, both as an aide memoire but also as a means of sharing with other pathologists, thus reducing problems of diagnostic drift. Rather than relying solely upon subjective opinions consideration should be given to use of quantitative morphometry for morphological features, where possible, for example in metabolic bone studies, and the accurate quantitation of protein expression detected by immunofluorescence.

3.6.3.7 Peer Review of histopathology

237. Peer review increases confidence in the accuracy of the histopathology findings from a study. Peer review by an independent pathologist is essential to ensure consistency of the histopathological findings in any studies but in particular those evaluated by more than one pathologist. The objectives of a formal histopathology peer review are several: 1) determine accuracy and consistency of nomenclature, i.e., survey for the presence of incorrectly diagnosed or inaccurately described treatment-related lesions, 2) determine completeness, i.e., survey for the presence of undiagnosed treatment-related lesions, 3) determine the appropriateness of the NOEL, NOAEL or other point of departure, e.g., BMDLx, by reviewing all target tissues and organs, and 4) review the correctness of the textual interpretations derived from those data. The methods employed may vary depending on the purpose of the peer review. For a routine peer review, tissues from a sufficient number of treated animals need to be evaluated to assure that significant lesions were not missed and that a "no effect level" can be verified.

238. A number of procedures can be used for peer review (Eighmy, 1996; Peters, 1996; The Society of Toxicologic Pathologists, 1991 and 1997). Peer reviews are generally included in the study protocol and are conducted prior to the issuance of the study report (Ward *et al*., 1995). It is important that the original plan for the review process include a joint review by the original and reviewing pathologist of the two sets of results to explain or resolve apparent differences, should they occur. The results of the peer review should be documented and archived and any differences of opinion resolved through consensus. This procedure is then part of the process that leads to finalizing diagnoses and interpretations.

239. If any differences cannot be resolved through a joint review of the data, then arbitration through a third pathologist should take place or the problem referred to a "pathology working group" (PWG) or to a panel of expert consultants for resolution. Contingency plans to allow for this should also be included in the original study plan. A Pathology Working Group (PWG) may be formed to review, revise and/or interpret diagnoses (Peters, 1996; The Society of Toxicologic Pathologists, 1991) and this should be done prior to finalising the report. The PWG should be composed of individuals having expertise both in the pathology of the test species and with the specific lesion in question. The PWG may include both the original and the reviewing pathologist but the diagnoses of both the original and the reviewing pathologist should not be known to the other PWG members. Prior to beginning work, a PWG chairman is appointed who directs the focus of the PWG and orchestrates

agreement of the criteria used to make the diagnoses of the lesions in question. For impartiality reasons, neither the original pathologist nor the reviewer pathologist should be nominated as chairman (Mann P.C., 1996). This process should be clearly defined and documented. The PWG may request additional sections from tissues to aid with the diagnostic process of the peer review. Following slide examination by the PWG in a blinded fashion and tabulation of the results; the findings of the PWG are compared to those of the original pathologist. When the consensus of the PWG is clearly different from that of the original pathologist, the diagnosis for an individual lesion should be changed. When there is a close split on a vote of a diagnosis, however, the diagnosis of the original pathologist should be allowed to stand. The records of the PWG should indicate the final diagnoses for each lesion and the degree of certainty of the diagnoses.

240. Regardless of its exact form, the peer review process should encourage direct interaction between the original and the reviewing pathologist and should result in the production of a single, scientifically robust pathology report that appropriately and accurately summarises the results and any uncertainties. The process should be constructive, meet the needs and objectives of the review and have an inbuilt procedure for resolving differences, should they arise.

241. Retrospective peer reviews by a single pathologist or a PWG may also be undertaken after the completion of a study. This type of peer review is necessary when unexpected issues are highlighted after a study is finalised and additional work is needed to clarify the issue. In these exceptional circumstances, where the conclusions of the study may change for example, it is necessary to take steps to ensure that the original report is marked to indicate that a revision of the final conclusions has occurred. A separate report or a report amendment may be appropriate and any changes must be documented as required by GLP.

3.6.3.8 Early Death Investigations

242. Guidance on the interpretation of early deaths is provided in OECD Guidance Document 35. In this Section further guidance on investigations that are valuable is provided.

243. In accordance with OECD Guidance Document 19 (GD 19), the earliest possible endpoints that are indicators of distress, severe pain, or impending death should be used as indications for humanely killing the animals prior to them reaching a moribund state or dying. GD 19 was developed largely for animal welfare reasons. However, this practice also has the advantage or reducing the number of animals that are found dead with the accompanying risk of loss of tissues from autolysis or cannibalism. It also avoids confusing direct effects of the test substance with secondary effects resulting from post-mortem or moribund changes.

244. Reasonable efforts should be made to determine the cause, or likely cause, of individual deaths or severe toxicity leading to euthanasia. It is important to identify causes unrelated to exposure to the test agent (e.g., acute or chronic infections, age or disease-related degenerative processes, anatomical abnormalities, mishandling or accident) from toxicity-induced effects. All in-life data, as well as the results of the post mortem, of euthanized or dead animals in a study, should be used in an attempt to make this distinction.

245. Assignment of the cause of death or of severe toxicity, leading to euthanasia is a component of the assessment of carcinogenicity studies in rodents, which is widely practised to aid statistical evaluation (Kodell *et al*., 1995). The US National Centre for Toxicological Research (NCTR) requests routine assignment of cause of death for all dead and moribund animals examined histologically in carcinogenicity studies. A determination as to whether a particular tumour under consideration was fatal to an animal or was simply an incidental finding at necropsy is important for determining the

type of information that an individual animal contributes to age-adjusted statistical tests for carcinogenic effects (Peto *et al*., 1995).

246. Organ weights from animals that die or are euthanized prior to scheduled necropsy are of little value in most cases to the overall study because of differences in nutritional status and tissue congestion and oedema secondary to the health status of the animal, and may not need to be recorded (Sellers *et al*., 2007). The absence of matched concurrent control data further hinders interpretation of organ weights from animals that die or are euthanized prior to termination of the study. Gross pathologic and/or histopathologic changes of the organ can however be useful in indicating a potential cause of death in the individual animal.

247. An experienced pathologist should, under appropriate conditions, be able to distinguish peri- and post-mortem changes in tissues from treatment-related effects on tissues. While cannibalism can inhibit accurate pathology evaluation it is almost always useful to fix and block tissues from all animals, even those found dead, since different tissues show different degrees of autolytic change in death and conclusions can often be made even when some time has elapsed between death and tissue sampling.

3.6.3.9 Pathology Final Conclusions

248. A final pathology/histopathology conclusion is reached when the following have been achieved: 1) histopathologic diagnosis and description of all findings for each animal on the study, 2) the diagnoses and descriptions reflecting the consensus of a peer review, if a peer review was done prospectively, 3) tabular summary data providing an accurate representation of the findings, 4) a NOEL, NOAEL, or other point of departure, e.g., BMDLx, for pathology/histopathological findings in the study, if appropriate. Final conclusions and diagnoses may often be the result of an iterative process in which the diagnosis is continually refined and modified until a final conclusion can be reached. This process may include taking additional samples for evaluation. Details of areas of disagreement and/or uncertainty that arose during the study and how these were resolved should be documented. The final pathology report must accurately and completely present the data and their interpretation.

249. Issues arise, particularly in carcinogenicity studies, as to whether findings are split into separate categories such as adenoma and carcinoma, grouped as benign and malignant, combined across multiple tumour sites and the inclusion or exclusion of metastatic tumours. Information on related metastases and its site(s) should be reported for the primary tumour. The general convention is that metastases from primary tumours should not be included separately in statistical analysis, except where no primary tumours could be identified. Benign and malignant lesions should always be presented separately. Combination may result in a loss of sensitivity if the substance tested accelerates the progression to malignancy, since the increase in malignant neoplasms is balanced by a decrease in benign neoplasms. However, benign and malignant neoplasms of the same cell type are sometimes combined because one is seen as the progression of the other. Therefore, whilst combined data may eventually be useful, the pathology narrative should be full and precise to prevent misinterpretation or doubts over the interpretation.

250. McConnell *et al.* (1986) recommended that the statistical analysis of each tumour type be carried out separately and, if it is considered scientifically defensible, further statistical analysis may be performed on the combined tumours (benign and malignant) of the same histogenic origin, even when those tumours are in different tissues: so for example, the incidence of lipoma & liposarcoma should be combined across all tissues, as should the overall incidence of vascular tumours and tumours of connective tissue types. More detailed guidance on combining neoplasms is given in

McConnell *et al.* (1986) and further guidance on statistical analysis is given in Chapter 4 of this Guidance Document.

251. The narrative of the final pathology report should address all the significant pathology findings, report any group differences and address the issue of disease progression. When the disease is progressive, then the identification in the pathological examination of the presence of the more severe form of a disease should automatically mean that the earlier (precursor) form is or has been present. Background pathology and other underlying conditions such as infections or the presence of incidental lesions in both control and treated animals should also be reported. There should be precise descriptions of any treatment-related findings and their likely importance. Clear reasons for concluding that any observations are unrelated to treatment or are not biologically or toxicologically significant should also be provided (Morton *et al*., 2006).

3.6.4 Supplementary Investigations

252. Previous knowledge of the effects of the substance may indicate that non-standard investigations would be of value in the interpretation of the results. Samples for such investigations should be taken and stored appropriately, provided that this does not jeopardise the basic requirements of the study. Such supplementary sampling would not thus invalidate the study as an OECD-compliant study according to the Test Guideline. Findings in the initial standard investigations may raise further questions that may be addressed by supplementary studies if appropriate stored samples are available.

3.6.5 Avoiding bias

253. Bias can arise if samples or observations in one group differ from those in another group for any reason other than treatment with the test substance. Therefore it is important that steps are taken to prevent the introduction of bias throughout the whole study and not restricted to the randomisation of animals to the treatment groups at the beginning of the study. Bias can be introduced whenever the operator or observer, equipment used or the timing of investigations are confounded with treatment group. The study design should minimise such effects by use of randomisation or a well balanced investigation scheme.

254. Randomisation should be done formally, not semi-subjectively. There are several methods for performing the randomisation process. The most commonly used is a computerised algorithm, but card assignment and the use of a random number table may also be used. Randomisation is normally used for the sequence of necropsy and can also be used for the sequence of other investigations. Similarly, if different operators or observers and equipment are used for the same investigation, randomisation should be used to prevent confounding with the treatment group. A well balanced investigation scheme can be used instead of randomisation for investigations.

255. Re-randomisation of samples is not necessary at each stage of the study and an initial randomisation or well balanced investigation scheme can be re-used throughout the study investigations. Slides are not normally randomised for histopathologic evaluation; slides from the control and highest dose groups are usually evaluated before other groups. The benefits and drawbacks of blinding or masking samples for histopathological evaluation are discussed below.

256. Subjective evaluations, such as clinical signs and histopathology are susceptible to bias if the evaluator is aware of the treatment group to which the individual animal or slide belongs. Blinded

or masked evaluation is carried out without prior knowledge of treatment group, which could include untreated or other control groups as a mechanism to minimize the introduction of observational bias. Blinded/masked evaluation is a more rigorous scientific approach to assessment.

257. For the evaluation of clinical signs, the absolute masking of the treatment group for individual animals is practically difficult as animals within one cage are usually from the same treatment group. However, blinding of the evaluator should be maintained to the extent possible without jeopardising other aspects of the study, particularly by increasing the risk of errors in dosing or animal identification.

258. Appropriate blinding procedures should be considered by the pathologist before, during, and after the microscopic evaluation of tissues. Blinding need not prevent the pathologist being provided with detailed information of other observations in the study, including animal-specific findings (see Section 3.6.3.4). The foremost objection to blinded slide evaluation in the initial histopathological examination is loss of knowledge of the range of normal that exists in known controls. This may be particularly pertinent for studies on new chemical entities or in new species, where little is known about potential treatment-related effects and their background frequency in control animals. Without this baseline, subtle differences between treated and control groups may be difficult to detect. Knowledge of the treatment group also allows the pathologist to assess the spectrum of related morphological changes and determine the most appropriate diagnostic terminology, including combining related diagnoses where indicated. However, these advantages must be weighed against the inevitable bias introduced by non-blinded evaluations. If an initial non-blinded examination is carried out to allow the recognition of treatment-related findings, subsequent masked evaluations of target tissues can be very helpful because it ensures an unbiased determination of a NOEL (no-observable-effect-level), or other point of departure.

259. Randomisation, well balanced investigation scheme and blinding or masking procedures employed to minimise bias in investigations should be described in the protocol and study report.

3.6.6 Data Archives

260. Archiving of documents and specimens generated during a non-clinical laboratory study is a fundamental Good Laboratory Practice (GLP) requirement. The records and material that should be archived as well as the characteristics and the organisation of archive facilities are addressed in the OECD series on Principles of Good Laboratory Practice No. 1 (1997) and the OECD GLP document 15 (2007). Each kind of material has its own individual requirements for proper storage and may have different retention times. Materials should be retained for the period specified by the appropriate authorities.

261. Raw data are defined in the OECD series on Principles of Good Laboratory Practice No. 1 (1997) as "all original test facility records and documentation, or verified copies thereof, which are the results of the original observations and activities in a study. Raw data also may include, for example, photographs, microfilm or microfiche copies, computer readable media, dictated observations, records data from automated instruments, or any other data storage medium that has been recognised as capable of providing secure storage of information for a time period as stated in section 10".

262. Any computerised system used to collect or store histopathology data should be GLP compliant. A variety of security measures will have been incorporated into each of these computer systems to ensure both data integrity throughout the process of histopathology data collection and reporting and to ensure that any changes made after the pathology contribution has been locked are

recorded, and that an audit trail for the changes, including who made them and when, is kept. These security controls include: limited user access, single and/or multiple password requirements, procedural controls and technical controls built into the systems. A critical requirement of an audit trail is a record of any change to the data.

REFERENCES

Abdo, K.M. and F.W. Kari (1997), "The Sensitivity of the NTP Bioassay for Carcinogen Hazard Evaluation can be Modulated by Dietary Restriction", *Exp Toxic Pathol,* 48: 129–137.

Bahnemann, R., M. Jacobs, E. Karbe, W. Kaufmann, G. Morawietz, T. Nolte and S. Rittinghausen (1995), "RITA – Registry of Industrial Toxicology Animal-data – Guides for Organ Sampling and Trimming Procedures in Rats", *Exp Toxic Pathol,* 47: 247–266.

Boorman G.A. and S.L. Eustis (1986), *The Pathology working group as a means for assuring pathology quality in toxicological studies*, In: Hoover B.K., Baldwin J.K., Uelner A.F., Whitmire C.E., Davies C.L. and Bristol D.W. (Eds) Managing Conduct and Data Quality of Toxicology Studies, pp. 271–275, Princeton Scientific Publishing Co., Princeton, N.J.

Bucci, T.J. (1991), *Evaluation of Altered Morphology*, In: Hascheck W.M. and Rousseaux C.G. (Eds) Handbook of Toxicologic Pathology, ch. 3, pp. 23–25, Academic Press, San Diego, Cal.

Cammermeyer, J. (1978), "Is the Solitary Dark Neuron a Manifestation of Postmortem Trauma to the Brain Inadequately Fixed by Perfusion?" *Histochemistry,* 56: 97-115.

Chhabra, R.S. and J.R. Fouts (1974), "Sex Differences in the Metabolism of Xenobiotics by Extrahepatic Tissues in Rats", *Drug Metab Dispos,* 6: 375–379.

Christian, M.S., A.M. Hoberman, M.D. Johnson, W.R. Brown and T.J. Bucci (1998), "Effect of Dietary Optimization on Growth, Survival, Tumor Incidences and Clinical Pathology Parameters in CD Sprague–Dawley and Fischer 344 Rats: a 104-week Study", *Drug Chem Toxicol,* 21: 97–117.

Darke, P.G.G., J.D. Donagura and D.F. Kelly (1996), *Color Atlas of Veterinary Cardiology,* Mosby, St Louis.

Dayton, P.G. and J.E. Sanders (1983), "Dose-dependent Pharmacokinetics: Emphasis on Phase I Metabolism", *Drug Metab Rev*, 14: 347–405.

Eastin, W.C. and R.W. Tennant (1998), *Carcinogenicity: Testing Predicting & Interpreting Chemical Effects,* In: Kitchin KT (Ed) pp. 395–410, Marcel Dekker Inc., New York.

Eastin, W.C. (1998), "The U.S. National Toxicology Program Evaluation of Transgenic Mice as Predictive Models for Identifying Carcinogens", *Environ Health Perspect*, 106(Sup. 1): 81–84.

Eighmy, J. J. (1996), "Study Pathologist Perspective of Pathology Peer Review." *Toxicol Pathol,* 24(5): 647–649.

Falkeholm, L., C.A. Grant, A. Magnusson and E. Möller (2001), "Xylene-free Method for Histological Preparation: a Multicentre Evaluation", *Lab Invest*, 81:1213-21.

Fix, A.S. and R.H. Garman (2000), "Practical Aspects of Neuropathology: A Technical Guide for Working with the Nervous System", *Toxicol Pathol* 28; 122.

Greaves, P. (2009), *General and Applied Toxicology*, 3rd Edition, Wiley Press, Editors Ballantyne, B, Marrs, T, Syversen, T. ISBN: 978-0-470-72327-2.

Hart, R.W., D.A. Newmann, R.T. Robertson (1995), *Dietary Restriction: Implications for the Design and Interpretation of Toxicity and Carcinogenicity Studies*, ILSI Press, Washington, DC.

Hausmann, R., H. Bock, T. Biermann and P. Betz (2004), "Influence of Lung Fixation Technique on the State of Alveolar Expansion - a Histomorphometrical Study", *Leg Med (Tokyo)* 6:61–65.

Hayward, J.J., B.S. Shane, K.R. Tindall and M.L. Cunningham (1995), "Differential *in vivo* Mutagenicity of the Carcinogen–noncarcinogen Pair 2,4- and 2,6-diaminotoluene", *Carcinogenesis,* 16: 2429–2433.

Hewitt, S.M., F.A. Lewis, Y Cao, R.C. Conrad, M. Cronin, K.D. Danenberg, T.J. Goralski, J.P. Langmore, R.G. Raja, P.M. Williams, J.F. Palma and J.A. Warrington (2008), "Tissue Handling and Specimen Preparation in Surgical Pathology: Issues Concerning the Recovery of Nucleic Acids from Formalin-fixed, Paraffin-embedded Tissue", *Arch Pathol Lab Med*, 132:1929-35.

House, D. E., E. Berman, J. C. Seely, and J. E. Simmons (1992), "Comparison of Open and Blind Histopathologic Evaluation of Hepatic Lesions" *Toxicol Lett,* 63: 127–133.

IARC (1992–1997) International Classification of Rodent Tumours, Part 1. The Rat (set of 10 fascicles), Mohr U (Ed), IARC Scientific Publications 122, Lyon.

Iatropoulos, M.J. (1984), "Editorial", *Toxicol Pathol,* 12(4): 305–306.

Keenan, K.P., G.C. Ballam, K.A. Soper, P. Laroque, J.B. Coleman and R. Dixit (1999), "Diet, Caloric Restriction, and the Rodent Bioassay", *Toxicol Sci,* 52 (Suppl): 24–34.

Keenan, K.P., J.B. Coleman, C.L. Mc Coy, C.-M. Hoe, K.A. Soper and P. Laroque (2000), "Chronic Nephropathy in ad libitum Overfed Sprague-Dawley Rats and its Early Attenuation by Increasing Degrees of Dietary (Caloric) Restriction to Cntrol Growth", *Toxicol Pathol*, 28: 788–798.

Keenan, K.P., P. Laroque, G.C. Ballam, K.A. Soper, R. Dixit, B.A. Mattson, S.P. Adams and J.B. Coleman (1996), "The Effects of Diet, ad libitum Overfeeding, and Moderate Dietary Restriction on Rodent Bioassay: The Uncontrolled Variable in Safety Assessment", *Toxicol Pathol*, 24: 757–768.

Kemi, M., K.P. Keenan, C. McCoy, C.-M. Hoe, K.A. Soper, G.C. Ballam and M.J. vanZwieten (2000), "The Relative Protective Effects of Moderate Dietary Restriction versus Dietary Modification on Spontaneous Cardiomyopathy in Male Sprague-Dawley Rats", *Toxicol Pathol*, 28: 285–296.

Kittel, B., C. Ruehl-Fehlert, G. Morawietz, J. Klapwijk, M.R. Elwell, B. Lenz, M.G. O'Sullivan, D.R. Roth and P.F. Wadsworth (2004), "Revised Guides for Organ Sampling and Trimming in Rats and Mice - Part 2 - A Joint Publication of the RITA and NACAD Groups", *Experimental and Toxicologic Pathology,* 55**,** 413-431.

Kodell, R. L., B.-N. Blackwell, T.J. Bucci, and D.L. Greenman (1995), "Cause-of-death Assignment at the National Center for Toxicological Research", *Toxicologic Pathology,* 23**,** 241-247.

Krinke, G.J., L.P.A. Perrin and R. Hess (1991), "Assessment of Toxicopathological Effects in Ageing Laboratory Rodents", *Arch Toxicol*, (Sup. 14) 43–49.

Latendresse, J.R., A.R. Warbrittion, H. Jonassen and D.M. Creasy (2002), "Fixation of Testes and Eyes Using a Modified Davidson's Fluid: Comparison with Bouin's Fluid and Conventional Davidson's Fluid", *Toxicologic Pathology*, 30, 524-533.

Levy, G. (1968), *Dose Dependent Effects in Pharmacokinetics*, In: Important Fundamental Principles in Drug Evaluation, pp. 141–172.

Mann, P.C. (1996), "Pathology Peer Review from the Perspective of an External Peer Review Pathologist" *Toxicol Pathol,* 24(5): 650–653.

McConnell, E.E., H.A. Solleveld, J.A. Swenberg and G.A. Boorman (1986), "Guidelines for Combining Neoplasms for Evaluation of Rodent Carcinogenicity Studies", *J. Natl.Cancer Inst,* 76 283-289.

McInnes, E. (2005), “Artefacts in Histopathology”, *Comp Clin Path*, 13: 100–108.

Mohr, U. and G. Morawietz (1995), “Recent Progress in the Field of Toxicologic Pathology in Europe and the Need for Standardisation”, *J Toxicol Pathol,* 8: 219–226.

Mohr, U., (1999), “RITA – Registry of Industrial Toxicology Animal-data: Optimization of carcinogenicity bioassays”, *Exp Toxic Pathol*, 51: 461–475 (editorial).

Morawietz, G., S. Rittinghausen, U. Mohr (1992), “RITA – Registry of Industrial Toxicology Animal-data – Progress of the Working Group”, *Exp Toxic Pathol*, 44: 301–309.

Morawietz, G., C. Ruehl-Fehlert, B. Kittel, A. Bube, K. Keane, S. Halm, A. Heuser andJ. Hellmann (2004), “Revised Guides for Organ Sampling and Trimming in Rats and Mice - Part 3 - A Joint Publication of the RITA and NACAD Groups”, *Experimental and Toxicologic Pathology,* 55**,** 433-449.

Morton, D., R.K. Kemp, S. Francke-Carroll, K. Jensen, J. McCartney, T.M. Monticello, R. Perry, O. Pulido, N. Roome, K. Schafer, R. Sellers and P.W. Snyder (2006), "Best Practices for Reporting Pathology Interpretations within GLP Toxicology Studies", *Toxicologic Pathology*, 34, 806-809.

Munro, I.C. (1977), “Considerations in Chronic Toxicity Testing: The Chemical, the Dose, the Design”, *J Environ Path Toxicol,* 1: 183–197.

Nassiri, M., S. Ramos, H. Zohourian, V. Vincek, A.R. Morales and M. Nadji (2008), “Preservation of Biomolecules in Breast Cancer Tissue by a Formalin-free Histology System”, *BMC Clinical Pathology*, 8:1doi:10.1186/1472-6890-8-1.

National Academy of Sciences/National Research Council (1983), *Risk Assessment in the Federal Government: Managing the Process,* National Academy Press, Washington, D.C.

National Academy of Sciences/National Research Council (1994), *Science and Judgement in Risk Assessment,* National Academy Press, Washington, D.C.

Newberne, P.M. and F.A. de la Iglesia (1985), “Editorial: Philosophy of Blind Slide Reading”, *Toxicol Pathol,* 13(4): 255.

OECD (2002), *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies*, Series on Testing and Assessment No. 35 and Series on Pesticides No. 14, OECD Paris.

OECD (2000) *Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation*, Series on Testing and Assessment No 19, OECD Paris.

OECD (2007), *Establishment and control of archives that operate in compliance with the principles of GLP*, Series on Principles of Good Laboratory Practice and Compliance Monitoring, Number 15, OECD Paris.

OECD (1998), *OECD Principles on Good Laboratory Practice*, Series on Principles of Good Laboratory Practice and Compliance Monitoring, Number 1, OECD Paris.

Pearson, G.R., E.F. Logan (1978), “The Rate of Development of Post-mortem Artefact in the Small Intestine of Neonatal Calves”, *Br J Exp Pathol*, 59:178–182.

Peters, T.S. (1996), “Pathology Peer Review—A Concept for Consideration” *Toxicol Pathol,* 24(5): 654–656.

Peto, R., M.C. Pike, N.E. Day, R.G. Gray, P.N. Lee, S. Parish, J. Peto, S. Richards and J. Wahrendorf (1980), *Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-term Animal Experiments*, In: Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal (IARC Monographs, Supplement 2), R. Montesano, H. Bertsch and L. Tomatis (eds). International Agency for Research on Cancer, Lyon, pp. 311-426.

Prasse, K., P. Hildebrandt and D. Dodd (1986), "Letter to the Editor" *Vet Pathol* 23: 540–541.

Rao, G.N. (1997), "New Nonpurified Diet (NTP-2000) for Rodents in the NationalToxicology Program's Toxicology and Carcinogenesis Studies", *J Nutr*, 127: 842S–846S.

Rao, G.N., R.W. Morris, J.C. Seely (2001), "Beneficial Effects at NTP-2000 Diet on Growth, Survival and Kidney and Heart Diseases of Fischer 344 Rats in Chronic Studies", *Toxicol Sci*, 63: 245-255.

Ruehl-Fehlert, C., B. Kittel, G. Morawietz, P. Deslex, C. Keenan, C.R. Mahrt, T. Nolte, M. Robinson, B.P. Stuart and U. Deschl (2003), "Revised Guides for Organ Sampling and Trimming in Rats and Mice - Part 1", *Experimental and Toxicologic Pathology,* 55**,** 91-106.

Sellers, R.S., D. Morton, B. Michael, N. Roome, J.K. Johnson, B.L. Yano, R. Perry and K. Schafer (2007), "Society of Toxicologic Pathology position paper: organ weight recommendations for toxicology studies", *Toxicol Pathol* 35 751-755.

Sharma R.P. and R.A. Coulombe, Jr (1996), *Pharmacokinetics and Risk Assessment*, In: A.M. Fan and L.W. Chang (Eds) Toxicology and Risk Assessment, ch. 7, pp. 81–99. Marcel Dekker Inc., New York, N.Y.

Society of Toxicologic Pathologists' Position Paper on Blinded Slide Reading (1986), *Toxicologic Pathology* 14(4): 493–494.

Society of Toxicologic Pathologists Position Paper: Documentation of Pathology Peer Review (1997), *Toxicol Pathol* 25(6): 655.

Society of Toxicologic Pathologists' Recommendations on Statistical Analysis of Rodent Carcinogenicity Studies (2002), *Toxicol Pathol* 30(3), 415–418.

Standardized System of Nomenclature and Diagnostic Criteria. *Guides For Toxicologic Pathology.* STP/ARP/AFIP, Washington, DC.

Summers, B.A., J.F. Cummings and A. De Lahunta (1995), *Veterinary Neuropathology*, Mosby Year Book, pp 34–35.

Thompson, S.W. and L.G. Luna (1978), "An Atlas of Artifacts Encountered in the Preparation of Microscopic Tissue Sections", Charles Louis Davis DVM Foundation, Gurnee.

Travlos, G.S., R.W. Morris, M.R. Elwell, A. Duke, S. Rosenblum and M.B. Thompson (1996), "Frequency and Relationships of Clinical Chemistry and Liver and Kidney Histopathology Findings in 13-week Toxicity Studies in Rats", *Toxicol,* 107: 17–29.

WHO (1987), *Principles for the Safety Assessment of Food Additives and Contaminants in Food,* IPCS Environmental Health Criteria 70, World Health Organisation, Geneva.

4. STATISTICAL AND DOSE RESPONSE ANALYSIS, INCLUDING BENCHMARK DOSE AND LINEAR EXTRAPOLATION, NOAELS AND NOELS, LOAELS AND LOELS

4.1 Preamble

263. This document is intended to provide guidance on the statistical issues associated with the design and analysis of chronic toxicity and carcinogenicity bioassays, the analysis and interpretation of resulting data and the use of these data in the identification of Benchmark doses, linear extrapolation and various no observed adverse effect levels (NOAELs) and lowest observed adverse effect levels (LOAELs) measures.

264. The fundamental point that this document aims to convey is:

"The statistical methods most appropriate for the analysis of results, given the experimental design and objectives, should be established before commencing the study" (OECD 2009, paragraph 9 of Test Guideline (TG) 451/453 and paragraph 8 of TG 452).

265. Statistical analysis of biological data is intertwined with the experimental design of studies so this document also includes discussion of issues related to study design. Consequently there is some overlap with topics discussed in other sections but this can allow relevant linkages and cross-references to be made to other sections of the document.

4.2 Introduction

266. The central concept of this document is that the experimental design represents the strategy for answering the question of interest and that the specific statistical analyses are tactical methods used to help answer the questions. Therefore, the statistical methods most appropriate for the analysis of the data collected should be established at the time of designing the experiment and before the study starts.

267. Many of the standard methods in long-term animal experimentation have been in use for a considerable time. There is, therefore, considerable experience of the statistical properties and strengths/weaknesses of these designs and in the integration of the biological importance of findings with the results of statistical analyses. There is, though, a need to ensure that studies are planned before beginning so as to optimize the use of the resources available and to avoid the time and costs associated with the unnecessary repeat of poorly designed or the conduct of duplicate studies. From an ethical perspective, there is a requirement to avoid the use of more animals than is absolutely necessary. Statistical analysis plans should be in place before the start of the study and staff with appropriate statistical expertise should be involved in all stages of the study, especially the design stage.

268. The principles of the formal study design developed over many years should continue to underpin the subsequent pragmatic interpretation often needed in the interpretation of data. It is important to ensure that the quality of studies does not decline because of familiarity with the methods. Long-term studies require a high standard of conduct to ensure unbiased statistical analysis. These requirements are well known and have not undergone any significant changes in recent years. (See paragraph 275 – Current statements on statistical methods.)

269. Much of this Chapter will concentrate on the statistical methods proposed in various documents and guidelines specifically developed for the analysis of chronic toxicity data. Particular

attention will be given to OECD Guidance Document N° 35 (2002) which made some general points about the strengths and weaknesses of statistical methods and listed some common statistical tests (reproduced and augmented here as Figure 1) and provided references to a number of sources.

270. It is important at this point to stress that there is no single approach to the statistical analysis of data. There are different schools of thought about statistical methodologies. These can raise fundamental, almost philosophical, issues about the role of statistical analysis. An example is the long-running debate between 'Frequentists' and 'Bayesians'. Statistical methods also continue to develop so that new and modified approaches may continue to be proposed. As a consequence there can be alternative approaches to those suggested in various guidance documents. Such methods may, in practice, satisfy the requirements of a regulatory authority but it is always recommended that such approaches are discussed in advance with the relevant regulatory authority.

4.3 Objectives

271. In the 1960s and 1970s the primary objective of a long-term rodent carcinogenicity bioassay was qualitative hazard identification (i.e., identification of chronic toxicity and the evaluation of the carcinogenic potential of a chemical administered to rodents for most of their lives).

272. The purpose of the long-term rodent carcinogenicity bioassay has, however, widened to extend to a number of other objectives (e.g., paragraph 6 of TG 451 or TG 453). These include objectives relating to hazard characterization, describing the dose-response relationship and the derivation of an estimate of a Point of Departure (POD) such as the Benchmark Dose (BMD) or a no observed adverse effect level (NOAEL) which can then be used to establish an acceptable level of human exposure.

273. As a consequence of multiple potential objectives study design can become a compromise with a trade-off in the ability to answer competing questions: hazard identification/characterization on the one hand and characterization of the dose-response on the other (see paragraph 278).

274. The objective of the chronic toxicity bioassay (TG 452) is to characterize the toxicological response of a substance in a mammalian species following prolonged and repeated exposure. Determining the carcinogenic potential is also the objective of the combined chronic toxicity/carcinogenicity study (TG 453), but is the primary objective of the carcinogenicity study (TG 451). In this guidance the term "long-term bioassay" will refer to both the carninogenicity and the chronic toxicity bioassay.

4.4 Current statements on statistical methods

275. OECD TGs 451, 452 and 453 contain sections relating to the statistical analysis of the data. All three guidelines state:

"When applicable, numerical results should be evaluated by an appropriate and generally acceptable statistical method. The statistical methods and the data to be analysed should be selected during the design of the study (paragraph 8 [TG452] or 9 [TG451] and [TG453]). Selection should make provision for survival adjustments, if needed."

276. In TGs 451 and 453, paragraph 9 states:

"The statistical methods most appropriate for the analysis of results, given the experimental design and objectives, should be established before commencing the study. Issues to consider include

whether the statistics should include adjustment for survival, analysis of cumulative tumour risks relative to survival duration, analysis of the time to tumour and analysis in the event of premature termination of one or more groups. Guidance on the appropriate statistical analyses and key references to internationally accepted statistical methods are given in a Guidance Document on the design and conduct of chronic toxicity and carcinogenicity studies , available on the OECD public website on Test Guidelines, and also in Guidance Document No.35 on the analysis and evaluation of chronic toxicity and carcinogenicity studies).”

This passage is the same as paragraph 8 in TG 452, with the exception of the second sentence which states:

“Issues to consider include whether the statistics should include adjustment for survival and analysis in the event of premature termination of one or more groups.”

277. Other organizations have made suggestions for the statistical methods to be used. For instance, the US EPA’s Guidelines for Carcinogen Risk Assessment (EPA, 2005) advises that:

"Statistical analysis of a long-term study should be performed for each tumor type separately. The incidence of benign and malignant lesions of the same cell type, usually within a single tissue or organ, are considered separately and are then combined when scientifically defensible (McConnell et al., 1986). Trend tests and pairwise comparison tests are the recommended tests for determining whether chance, rather than a treatment-related effect, is a plausible explanation for an apparent increase in tumor incidence. A trend test such as the Cochran-Armitage test (Snedecor & Cochran, 1967) asks whether the results in all dose groups together increase as the dose increases. A pairwise comparison test such as the Fisher exact test (Fisher, 1950) asks whether an incidence in one dose group is increased over the control group. By convention, for both tests a statistically significant comparison is one for which p is less than 0.05 that the increased incidence is due to chance. Significance in either kind of test is sufficient to reject the hypothesis that chance accounts for the result. A statistically significant response may or may not be biologically significant and vice versa. The selection of a significance level is a policy choice based on a trade-off between the risks of false positives and false negatives. A significance level of greater or less than 5% (the most common significance level) is examined to see if it confirms other scientific information. When the assessment departs from a simple 5% level, this should be highlighted in the risk characterization. A two-tailed test or a one-tailed test may be used. In either case a rationale is provided. "

4.5 Study designs

278. Strategic issues relevant to study design include the number of doses, their spacing, the choice of the top dose, the group sizes, the length of study and the choice of control groups. In particular, the choice of the number of dose levels and the dose spacing is crucial to achieving the objectives of the study (e.g., hazard identification or dose-response/risk assessment) and is important for subsequent statistical analysis. An important consideration in the design of studies is the statistical power of the study. This is the probability of detecting an effect of a certain size as statistically significant if it really exists. Most toxicological studies evolved from designs developed before formal power calculations were incorporated in to experimental design.

279. OECD TGs 451, 452 and 453 have a core design consisting of three treatment groups (with different dose levels) and one or more negative control groups for each sex. Each group should be at least 50 animals of each sex for TG 451 and 453 and at least 20 animals for each group for TG 452.

280. The OECD TG 453 recommends, for the chronic phase of the study, at least three dose groups and a control group, each group containing at least 10 males and 10 females per group.

281. There are different strategies for the allocation of resources (i.e., the animals) to the groups in the design depending upon the objective of a study. These strategies could range from the equal allocation between a single negative control and high dose group to maximize the power to detect a difference (hazard identification) to the allocation of single animals to a large number of different doses across the whole dose range. The first case would require an ANOVA-style analysis while the latter would use a regression analysis with a test for the lack of fit of the dose-response relationship. Intermediate designs could be different numbers of animals allocated to each of a number of different dose groups. This design can be analysed by the ANOVA methodology where the between group comparison can be broken down into linear, quadratic, other components and a lack of fit component. This above strategy reflects the continuum that exists between the ANOVA and regression modelling approaches.

282. There has been debate over whether a 4 group 50 animal/sex/group design should be replaced with an 8 group 25 animals/sex/group design. Unpublished work examined the power of the different designs, using three different scenarios to detect a linear trend in the proportions together with other dose selection issues using the nQuery Advisor software. These calculations made no assumptions about differential survival. This analysis showed that the power of the 8 group design was between 5 and 22% lower than the 4 group design. To achieve comparable power with the 8 group design, the 8 group sample sizes would need to increase by about 35 to 40%. However, both TG 451 and 453 recommend 50 animals/sex/group so that the alternative 8 group design is not recommended.

4.6 Control groups and length of study

283. The control groups can be either untreated or a vehicle control group. The animals in these groups are expected to be treated in an identical fashion to those in the test groups. A discussion of the implications for the statistical analysis of the inclusion of more than one control group can be found in the section 4.21 'Use of control data and dual control groups' (paragraphs 393-397). A control group of pair-fed animals may be included if the palatability of the substance administered in the diet is of concern (i.e., a continuous reduction of 20% of more in food intake).

284. All animals should be treated identically throughout the length of the study (see section 3.3.3 on study duration). At the end of the study a full detailed gross necropsy should be carried out on all the animals from the control and test groups.

285. Planned interim kills may be part of the design. It is suggested in the OECD TG 453 that a group of 10 male and 10 females per group (reduced from the original 20 per sex per group) should be included and that the terminal kill of these animals could act as an interim kill for the main carcinogenicity study. The justification for this reduction is that more information is available from the animals in the carcinogenicity phase. The proposed statistical analysis of the results from the interim kill should form part of the statistical analysis plan of the whole study to avoid introducing biases but the data should not be combined with the data from the main study.

286. The investigator should consider the implications of including such a small interim kill sub-group on the power of the study. Animals, such as those from the chronic portion of TG 453, should be allocated to the treatment group before starting the study based upon some randomisation process (see Bannasch *et al*., 1986). Identifying these animals before the start of the study could, however, lead to them experiencing slightly different test environments such as this sub-study being maintained in a different room under slightly different conditions which might introduce biases into the statistical analysis. Therefore, precautions such as ensuring all animals are exposed to the same environmental factors and animal handlers being blind to which animals will be in the interim kill group should be instigated.

4.7 Purpose of statistical analyses

287. The objective of the statistical analyses of the data generated in long-term toxicity tests is to assess whether exposure to the test chemical is associated with adverse effects such as, for instance, an increased tumour incidence. The results of these analyses will then help achieve the other objectives of the study outlined in paragraph 272.

288. Statistical analyses can address this aim in two ways. On the one hand, the objective may be to test a hypothesis that one or more treated group is different from the concurrent group; alternatively, the objective may be to estimate the size of an effect in a comparison between groups and provide some indication of the precision or confidence that can be ascribed to that estimate.

289. As mentioned previously there are different 'schools of thought' about the statistical analysis of data. Much of the work in toxicology has been carried out based upon a traditional frequentist approach particularly around the concept of hypothesis testing. While recognizing that alternative viewpoints exist and this is a controversial area, most of the emphasis in this document will be on the traditional approaches.

290. There is a need to remain aware of the distinction between statistical significance and biological importance. Statistical analysis involves more than the reporting of the statistical significance of a hypothesis test. This may be one, often small component, of the much larger component of the design, analysis and interpretation of an experiment with statistical analysis being a part of the interpretation of the biological importance, not an alternative. Many statisticians argue against the reporting of significance levels arguing instead that the emphasis should be on emphasising the size of effects and the confidence in them. This avoids the problem of a small biologically unimportant effect being declared statistically significant and the artificiality of trying to dichotomise a result into a positive or negative finding on the basis of a P-value of, for instance, either 0.051 or 0.049. When reporting the results of significance tests, precise P-values (e.g., P=0.051) should be reported rather than referring to specific critical values.

291. The concept of statistical significance is an important component of the hypothesis testing approach. In a test of a null hypothesis the P value is a measure of how likely a result that has been obtained, or one more extreme, might have arisen if there were no difference between, for example, the two groups. The P value is dependent upon a number of factors: endpoint, variability, sample size, experimental design and statistical method. Conventionally, certain critical values ($P < 0.05$, $P < 0.01$ and $P < 0.001$) denote specific levels of statistical significance although many statisticians dislike this approach. Denoting something as statistically significant does not mean it is biologically important. (The use of the term biologically or clinically important is an attempt to avoid misunderstandings as the word 'significant' has a specific and precise meaning for statisticians.)

292. Similarly, declaring a result non-significant (often designated as $P>0.05$ or NS, again a nomenclature not favoured by statisticians) should not be interpreted as meaning the effect is not biologically important or that the null hypothesis is correct. Rather it means that there is not sufficient evidence to reject the null hypothesis.

293. Selecting an appropriate statistical method to analyse data is dependent on both the study purpose and the type of data. The specific tests applied will have different results because they are testing different hypotheses. A trend test will be testing whether there is a linear trend with a slope greater than zero; a pairwise comparison will be comparing whether a treated group is significantly different from the controls. In general, testing a trend which is a more specific hypothesis has greater power than a pair-wise comparison. It is also a single test compared with the 3 pair-wise comparisons between the dose groups and the negative controls. This introduces the concept of the use of corrections for multiple comparisons. Corrections are sometimes used to address concerns that when a

large number of comparisons (e.g., between pairs of treatments) are made that there is a risk of Type 1 errors. (A Type I error is the risk of wrongly rejecting the null hypothesis in a statistical test when, in fact, it is true and thus declaring results significant when they are not).

Parametric versus non-parametric methods

294. Deciding whether to use parametric or non-parametric methods to analyze data can be an issue, for each basic parametric test has a non-parametric counterpart (e.g., a pairwise comparison via Student's t-test or Mann-Whitney U-test; multigroup mean comparison via 1-way ANOVA or Kruskal-Wallis; trend test via a linear dose-response or Jonkheere-Terpstra test.)

295. Non-parametric tests have an advantage when some of the assumptions, particularly normality, underlying parametric tests are violated. They then give exact and accurate probabilities regardless of the shape of the distribution the data were randomly sampled from. However, while non-parametric tests may be distribution free, they are not assumption free so are as vulnerable, if not more so, to differences in the distributions between the groups. Non-parametric tests aim to ensure that correct Type I errors are derived but are less suitable for more complex designs, estimation and model fitting. When the assumptions underlying parametric methods are met their non-parametric equivalents have lower power and thus waste data. In the case of small sample sizes (e.g., 4 or 5 experimental units per group) comparisons using non-parametric tests have low power to detect even quite large treatment effects.

Limitations of statistical analysis

296. It should always be appreciated that a statistical analysis has its limitations. Statistical analysis cannot rescue poor data resulting from a flawed design or a poorly conducted study. Good experimental design, again the 'strategy' (paragraph 266), is the critical role of statistical analysis in a study. An appropriate data analysis will follow directly from a correct experimental design (including the selection of statistical methods to be applied) and implementation.

4.8 Types of data: Qualitative and Quantitative endpoints

297. Different types of data are collected in the course of a long-term bioassay. Endpoints can be qualitative or quantitative. The power associated with the detection of biologically important effects can be very different between a qualitative and a quantitative endpoint.

298. Qualitative data can be binary, categorical or ordinal. Examples of binary data are where the classification can take one of two (binary) forms: an animal can be dead or alive or have a tumour or not.

299. There are important issues about how pathological findings are described (see section 3.6.3.9). Sometimes these can be considered categorical (no ordering) or ordinal (where there is some ranking or ordering of the types). Issues arise as to whether findings are split into separate categories. This can alter the background incidence of the tumour types which affects power considerations. Splitting into a number of different separate categories of tumours also raises multiple comparison issues. Benign and malignant tumours should be analysed separately (McConnell *et al.,* 1986; US EPA, 2005) and, if it is considered scientifically defensible, further analysed using the combined incidences.

300. Further guidance on reporting of tumours and other histopathological lesions is provided in Section 3.6.3.9. Any final decision on the datasets to be analyzed should be agreed with the study pathologist.

301. Pathology data are categorical but can be converted into semi-qualitative or ordinal data. For instance, histopathology grading can be separated into categories of gradation from no effect, through mild to more severe. The assumption is that there is increasing severity but there is no assumption that the differences between classes are on a linear scale. (For instance, a change from a severity Category I to Category II may not be directly comparable to a change from Category II to Category III). Dichotomisation or other categorisation of continuous or ordinal variables may sometime be desirable, but categorisation will result in a loss of information.

302. A considerable amount of quantitative data is collected during the course of a long-term bioassay. Much of this is continuous data such as body and organ weights; clinical chemistry and haematological data (e.g., white blood cell counts) are also collected in chronic toxicity studies. In the case of carcinogenicity studies, the length of time either to the death or the identification of a tumour is a quantitative measure (used in, some cases, as a surrogate measure of the time until a tumour arises). There can also be quasi-continuous data where, although the data are discrete counts (e.g., the numbers of various types of blood cells), these counts are such large numbers that they can be analysed as if they are continuous data,

303. The specific statistical methods or tests used to analyse qualitative and quantitative endpoints are different. These have been represented as various times in a decision tree format with different flowcharts. An example developed by the OECD (2002, Guidance Note 35) previously will be discussed below (See pargraph 309 onwards). It is important to appreciate that while the statistical methodologies and algorithms differ, the underlying statistical concepts associated with the interpretation of the tests are basically the same. However, the power associated with the different endpoints within the context of the same basic experimental design can be very different.

4.9 Sample size and power considerations

304. The power of a study is the probability of detecting a true effect of a specific size or larger using a particular statistical test at a specific probability level. The power is (1-β) where β is the Type II error associated with a hypothesis test. (The Type II error is the probability of wrongly accepting the null hypothesis as true when it is actually false while the Type I error is the probability of rejecting the null hypothesis when it is actually true.)

305. The power of a study for a qualitative trait depends upon 5 factors: the sample size (n), the significance level (α), whether the test is one- or two-sided, the size of effect of interest (d = q-p) where q is the incidence in the treated group and p is the control incidence and the type of statistical test. In the case of quantitative data; the proportions are replaced by the size of effect of interest and a measure of the inter-individual variability such as the standard deviation. Numerous software packages and programs are available for carrying out these calculations.

306. The OECD Test Guidelines indicate the appropriate sample sizes for each group. In the carcinogenicity study, the sample size is usually at least 50 animals of each sex at each dose level. This group size reflects a trade-off between the statistical power of the design and economic practicalities of the design. In practice, the carcinogenicity study has low power in the sense that treatment effects that might be considered biologically important cannot be detected routinely as statistically significant.

307. It is recognized that the power of the study can only be increased modestly by increasing sample sizes. Similarly, allocating animals differently between the test groups can increase the statistical power of detecting, for instance, an effect in a particular dose group (Portier & Hoel, 1983, 1984).

308. A common feature is that the power associated with qualitative data is less than that associated with quantitative data. The carcinogenicity study design, for instance, has low power for comparisons focussed on the low dose group and is only able consistently to detect large increases over the negative control incidence in tumour incidence with the power being reduced (further) if there is a high control incidence (Haseman, 1984).

4.10 Selection of Statistical Methods using Flowcharts

309. Flowcharts which provide a decision tree for the choice of statistical tests have been developed and used extensively in the analysis of statistical data in the biological sciences. There are obvious practical advantages in having a set of standard methodologies with the choice of particular methods made at key decision points based upon data. A number of examples can be found in textbooks. Gad (2006) for example has developed some for the analysis of toxicological data. The OECD produced a flowchart (Guidance Note 35, OECD 2002) in a previous document which is reproduced here (Figure 1).

310. The choice of the statistical method to use is based upon whether the data are qualitative or quantitative and upon the assumptions required by the test being met. The choice of route through the flow chart is based upon the results of answers to queries higher in the chart. For example, in the event of a test for non-normality of the data being statistically significant a non-parametric test may be chosen in preference to a parametric one. This methodology speeds the analysis and reduces the amount of valuable time a statistician spends on an analysis.

311. Gad (2000), for instance, suggested that statistical analysis of endpoints such as body and organ weight data are "universally best analysed by ANOVA followed, if called for, by a post hoc test." He suggests Bartlett's test is performed first to ensure homogeneity of variances. With smaller sample sizes he suggests that Kruskal-Wallis test may be more appropriate. In the case of clinical chemistry he points to the limitations of univariate analyses when the analysis must consider a battery of biochemical parameters to determine the overall effect.

312. Critics point out that, although there are efficiency gains in the application of flow charts, there is a 'de-skilling' of the task, an over-emphasis on significance testing for decision making and vulnerability to artefactual results. There is also the methodological problem of the use of a multiple testing procedure where one hypothesis test is used to choose another test which can complicate quantifying the true probability values associated with various comparisons.

313. Some concern has been expressed over whether tests for normality or for heterogeneity of variances are over-sensitive and, as a consequence, unnecessarily rule out the use of robust statistical methods such as the analysis of variance and, thus, potentially reduce the power of the design.

314. In the case of tests for normality, the null hypothesis is that the data are normally distributed. In practice, this can mean that a small sample which may deviate from a normal distribution may fail to 'trigger' a significant result, while a large sample with a slight deviation from a normal distribution may be considered significant and lead to a switch in the subsequent statistical test. Such test behaviour is counter-intuitive to what in practice is required in the selection of statistical tests and the switch between different statistical tests should not be automatic.

315. Treatment effects can result in heterogeneity in response between individuals leading to violations to the assumptions such as non-normality and heterogeneity of variances. Attempts to remove these violations by the use of transformations such as the logarithmic transformation are not always successful. One approach is then to apply both parametric and non-parametric tests to such data. It is important to realise that different statistical methods will produce different results (i.e.,

probability values) when they are applied to the same data sets. Concordant results then give increased confidence in the findings while discordant results 'flag' up anomalous data.

316. Other decision rules based upon trying to find an optimal transformation can lead to a very heterogeneous set of analyses based upon decision rules which lead on to another different transformation. The US FDA Redbook (FDA 2007), for instance, indicated that unnecessary transformation should be avoided as should the use of a mixture of parametric and non-parametric methods on the same endpoint.

317. The use of decision rule-based statistical analysis differs from the more current way of carrying out analyses using statistical models. In this there is an iterative process of fitting and testing models and checking assumptions.

318. Visual representation of data is also an important aspect of the analysis, relying on inspection of the data for outliers, trends, goodness of fit and checks of assumptions. Care should, therefore, be taken in carrying out statistical analyses using flowcharts. The usual considerations for the interpretation of statistical analyses should always be kept in mind.

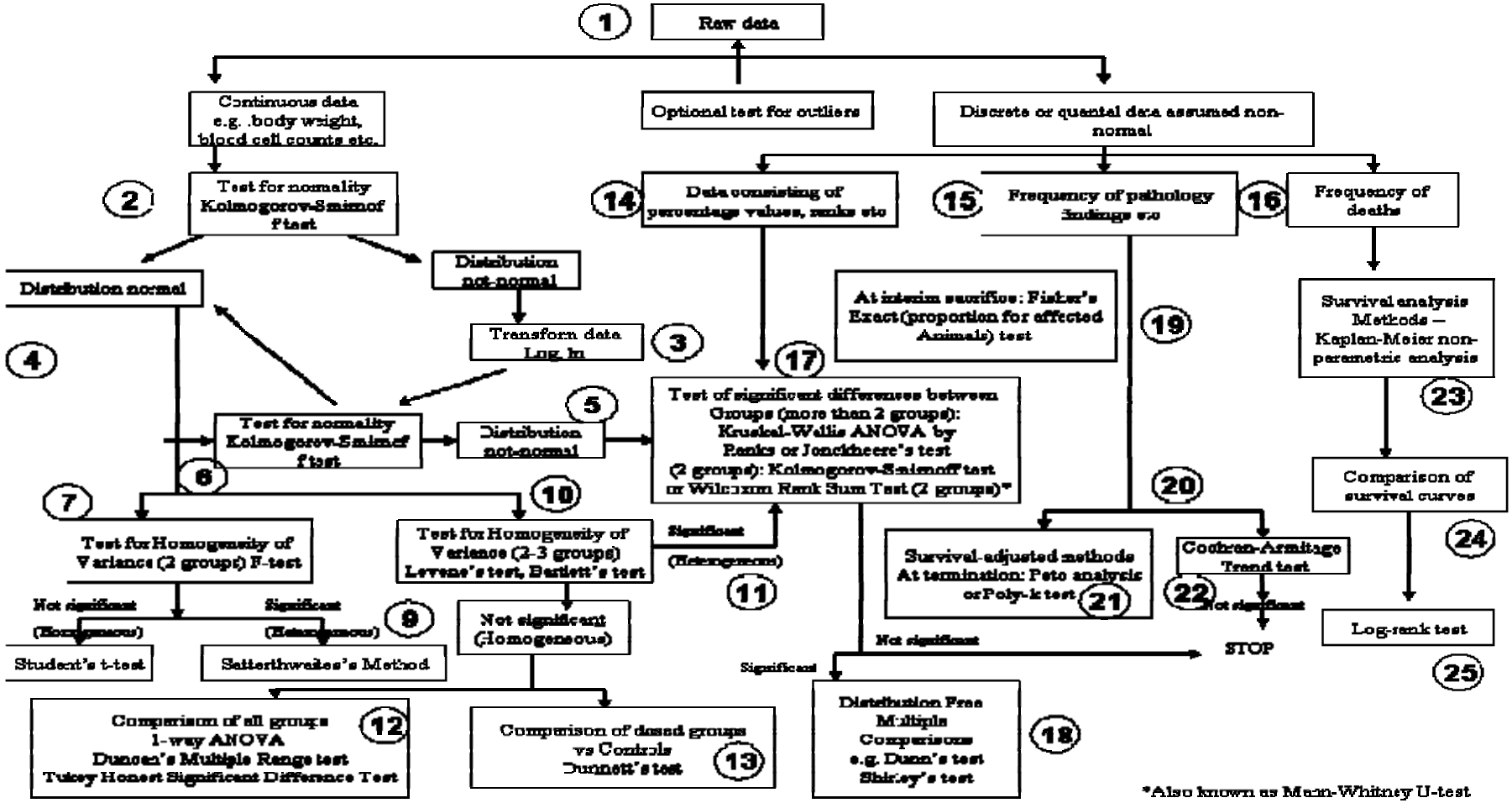
4.11 Description of the OECD flowchart

319. In this section Fig 1, a slightly modified version of the OECD (Guidance Note No 35; OECD, 2002, Fig 1) flowchart, will be worked through to illustrate the points that arise in its use. This will then be followed with some of the issues that arise if statistical analyses are totally dependent upon such an algorithmic approach. The objective, here, is to provide a brief overview of the methods used. The individual tests are described briefly in the glossary. This flowchart is similar to an approach used by the US National Toxicology Program (NTP).

320. For simplicity the description of the flowchart will be by working from left to right. This is for convenience and there is no priority implicit in this ordering. When an option is reached, the left choice initially will be described. Once the end of the 'tree' has been reached the description will move back to the previous node/decision point and continue down that. The same 'zig zag' procedure will be carried until the tests at the far right of the flowchart are reached.

321. In general, (but not precisely) the boxes at the top are concerned with aspects of the nature of the data. Moving down there are tests of the assumptions underlying the methods followed by 'omnibus' tests (which test for overall differences between groups irrespective of the specific design). These are followed by tests of whether there is a linear trend across the treatment groups and/or between treatment groups and the negative control group. In some cases there is a circular/interactive path when, for instance, following rejection of the assumption of normality the data are transformed to logarithms (natural or to the base 10) and the test for normality made again. The numbers in brackets refer to the circled numbers representing specific points on the flowchart. The assumption is that a decision is made to choose one or other route if the P value associated with the test statistic is <0.05. (Note that the use of hypothesis tests, the specific choice of P values, the choice of parametric or non-parametric and the use of multiple testing procedures are controversial issues and the particular routes outlined in this particular flowchart would not necessarily be agreed on by all statisticians.) Note that some of the statistical methods, particularly those related to the analysis of tumour data are discussed in more detail in paragraph 346 onwards.

Figure 1 A statistical decision tree, summarising common statistical procedures used for the analysis of data in long-term toxicology studies.



322. The top level of the flowchart (1) relates to a check of the data for overall quality and the identification of the type of data, whether quantitative continuous or qualitative or discrete. As part of the data checking an optional test for outliers such as the Dixon & Massey test is suggested. Moving to the methods included for continuous data, a test for the assumption of normality (either the Kolmogorov-Smirnov test or the Shapiro-Wilk) test is identified (2). If the test is significant indicating that the data are not normally distributed, the option is that the data are logarithmically transformed (3) and the test for normality carried out again as well as perhaps testing for outliers using the Extreme Studentized Deviate statistic (4). If neither transformation or identifying outliers results in normality the suggestion is to assume the distribution of the data is not normal and move (5) to the use of non-parametric methods (in the centre of the flowchart). If the data are assumed to be normal then a further test for homogeneity of variances (6) is suggested. In the case of a two group comparison this is an F-test (7). If the F test is not significant then the two groups are analysed by a standard Student's t-test (8); if the F test is significant then the comparison is by the modified t-test using Satterthwaite's method for unequal variances (9). Returning to point (6): either Levene's test or Bartlett's test are used to test for homogeneity of variances (10) when there are three or more groups. If the variances are considered heterogeneous the flowchart directs the analysis to non-parametric methods (11). If the variances are considered homogeneous then the comparison of all groups is suggested to be firstly by a one-way analysis of variance (1-way ANOVA) (12) then followed by a decision whether to proceed to a multiple comparison procedure such as Duncan's multiple range test or Tukey's Honest Significant Difference test. A number of other multiple comparison methods are available such as the Williams multiple comparison test (Williams, 1971). In the case of pair-wise comparisons between the control and the dosed groups, the flowchart suggests Dunnett's test (13) following the one-way analysis of variance.

323. Returning to the qualitative data (1) a distinction is made between data which are ranked or are discrete counts of an endpoint within the experimental unit (14) and those which are pathology findings (15) and those related to death or survival (16) The pathology findings and number of deaths may be summarized as frequencies or percentages for the different experimental groups. It is suggested that ranked data, together with quantitative data which were either determined to be non-normally distributed (5) or have heterogeneous variances (11) are analysed by non-parametric methods (17). The suggested methods for comparisons are the Kruskal-Wallis test for comparisons between the groups and Jonckheere's test for a trend in the data. Methods identified for comparisons between the control and test groups are the Kolmogorov-Smirnoff test and the Wilcoxon Rank Sum Test (which is equivalent to the Mann-Whitney U-test). If these tests are significant then further testing using distribution free multiple comparison tests can be carried out using tests such as Dunn's or Shirley's tests (18). Returning to pathology findings (15), if an interim kill is carried out comparisons between the proportions of animals with pathological findings in a treated group can be compared with the proportion in the control group using Fisher's exact test (19). An overall test of differences in proportions between the groups can be carried out using the chi-square test of heterogeneity outlined in paragraph 341. At the end of the study a choice is made (20) between tests which take into account information on how long the animals lived without a tumour (21) and those that do not such as the Cochran-Armitage test for a trend in proportions (22). The survival adjusted tests (21) are the Peto analysis (which requires information about whether the tumour is 'incidental' or 'fatal') and the poly-k test which does not need this information. (These methods are described in more detail in the survival adjusted analyses section of this Chapter, paragraph 346 onwards.)

324. Returning to qualitative data (1) and to data on survival/death (16), the flowchart identifies survival analysis approaches such as the Kaplan-Meier non-parametric methods (23) followed by comparison of the graphs of the survival curves (24) followed by analysis using the log rank test (25).

4.12 Intercurrent mortality

325. Inter-current mortality is death that arises during the course of a study from anything other than a tumour. Chronic studies can last up to two years. Animals can and do die in the course of the studies for a variety of reasons both related and unrelated to the treatment they have received. Inter-current mortality is different from the planned interim sacrifices or kills specifically included in some designs. Inter-current mortality complicates the statistical analysis of comparisons between test groups. For instance, older animals are more likely to develop a tumour than younger ones. The risk of getting a tumour and of dying because of a tumour increases with age. Consequently, the probability that an animal that dies unexpectedly during the course of a study also has a tumour will depend upon the animal's age at death. The test chemical may also affect the survival of different groups by causing either more deaths (through non-tumour related toxicity) or fewer deaths (by, for instance, reducing food intake and making the animals less susceptible to obesity-related morbidity and mortality). Peto *et al*. (1980) point out that comparisons based upon crude tumour rates in the presence of differential mortality can produce serious errors in the analysis.

326. A simple statistical analysis which does not account for inter-current mortality (described in paragraph 341) can underestimate the carcinogenic effects if the treatment decreases survival. Conversely, if the treatment increases survival then the tests may overestimate the carcinogenic effects. Failure to take intercurrent mortality into effect can, therefore, produce serious biases in the interpretation of results. Peto *et al.* (1980) state "the effects of differences in longevity on numbers of tumour-bearing animals can be very substantial, and so, whether or not they (the effects) appear to be, they should routinely be corrected when presenting experimental results" They argued that to avoid this problem from occurring, adjustments are needed for differences in survival between the groups and this correction should be routinely used.

327. Inter-current mortality, if not related to the tumours, is a serious problem for the interpretation of chronic studies because a high number of animals in the groups need to survive for a successful analysis (see paragraphs 158-159 for a more detailed discussion).

328. There are occasions when inter-current mortality may not be a major problem. However, it is important that any organization carrying out such an analysis should contact the relevant regulatory authority for advice in the event of problems with survival so as to try to ensure that the optimum amount of information is obtained from any study which has to be stopped early.

4.13 Cause of death (COD) and Context of observation (COO)

329. The context of observation (COO) relates to whether the tumour is considered fatal or incidental. Cause of death (COD) relates to the specific cause of death of the animal.

330. Information on the cause of death (COD) of the animal and the 'context of observation' of a tumour (COO) (Peto, 1974)) may be important for subsequent statistical analysis of a study (step 21 in the flowchart). Peto *et al.* (1980) argue that a distinction needs to be made between whether a tumour is called 'fatal' or 'incidental' for a correct statistical analysis to be carried out. The 'context of observation' (COO) distinguishes between whether a 'fatal' tumour is one that is considered to have caused the death of the animal; an 'incidental' tumour is found when the animal died from an unconnected reason or was found at the terminal kill at the end of the experiment. Distinguishing between these two COOs can be both difficult and controversial. However, this distinction is a critical feature of the 'Peto' analysis described in the 1980 IARC monograph (Peto *et al*., 1980).

331. Complications arise because this dichotomy is not always simple. Information on the COO may not be available because it is not provided by the pathologist or is unreliable because it is based upon assumptions about lethality which may be contentious.

332. A pathologist may not be able to determine the COD of an animal. Some pathologists have argued that it is difficult retrospectively to diagnose accurately if a tumour is the true cause of death of an animal. There may not be a single factor which solely determines the death; there may be multiple causes of death with the presence of a tumour being just one of them. Alternatively, more than one tumour may contribute to the cause of death.

333. Peto *et al.*'s (1980) approach to this problem was to suggest four categories of tumour COO - (1) probably fatal, (2) possibly fatal, (3) possibly incidental and (4) probably incidental - with the suggestion of combining the categories in different ways using different cut-offs to produce a binary endpoint: e.g., combining 1+2 versus 3+4; 1+2+3 versus 4 and 1 versus 2+3+4 and analyzing each combination as a form of sensitivity analysis. This has proved controversial because it is argued that it is a device to subsequently reduce the categories back to two: 'fatal' and 'incidental'.

334. Lee & Fry (cited in Lee *et al.*, 2002) point out that the 'fatal' definition is often misunderstood by pathologists. The key issue is that a 'fatal' tumour is wholly responsible for the death of the animal not that the tumour had the potential to kill the animal at some time in the future. Lee and Haseman (cited in Lee *et al*., 2002) also disagree on how easy it is for the pathologist to make a distinction between the COOs. Lee quoted Peto's example of the high proportion of definitive diagnoses in the BIBRA nitrosamine study where 94% of over 4500 tumours could be classified as either "definitely incidental" or "definitely fatal," even though the pathologists had initially expressed reservations about whether such classifications could be made reliably. Haseman (cited in Lee *et al.,* 2002) argues that there were instances where it was difficult to make the distinction and when made it was often incorrect. There was also a tendency for the over-designation of fatal tumours (Abu *et al*., 2001; Kodell *et al.*, 1982). Haseman argues that pathologists should be allowed the freedom to make their own judgements. Soper and Kodell (cited in Lee *et al.*, 2002) argue for a more objective classification based upon the large historical data base available.

4.14 Time to tumour onset

335. The time until an event such as death occurs can be analyzed by the statistical technique of survival analysis. There is much widely available statistical software that can be used to carry out the standard analyses. Such methods are used to analyze the survival data from the carcinogenicity study (see paragraph 346 onwards).

336. Analysis of time to tumour data, however, is less straightforward. Ideally, the time when a tumour first occurred is needed so that these times can be compared between treatment groups. There are some tumours which can be observed during routine observation of an animal such as some skin and some palpable mammary tumours. In practice, the time, when a tumour is identified is an arbitrary but objective endpoint such as recognizing when the tumour reaches a certain size. These tumours are called 'mortality independent' and can be analysed by standard life-table survival analysis methods.

337. However, these are the exception. Most tumours are internal (and, therefore, hidden or 'occult') and will usually be detected only during a post mortem examination of the animal. The specific time (the age of the animal or the tumour onset time) when the tumour initially arose is consequently unknown. It cannot simply be replaced by the time when the tumour was first identified as an 'occult' tumour (in other words the time of identification of an incidental tumour is not a surrogate for the time of onset). This lack of information complicates analysis. Statistical methods

which take into account the time until an event occurs (such as survival analysis methods) need to make a number of assumptions for analyses of such 'occult' tumours.

338. Dinse (1994), for instance stated: "without direct observations of the tumour onset times, the desired survival adjustment usually is accomplished by making assumptions concerning tumour lethality, cause of death, multiple sacrifices or parametric models" (Dinse, 1994). Approaches to get around this problem are to model and impute the onset times, to include extra data (context of observation) or to make more assumptions (poly-k test)" (see paragraph 361).

339. One suggested approach to the problem is to increase the sample size and have planned interim kills/'sacrifices'. In practice, this is rarely done and there are no universal guidelines for the analysis of such studies.

4.15 Standard (simple) statistical analysis of qualitative data

340. Tests for pair-wise comparisons and trends are recommended for the statistical analysis of tumour data to test for treatment-related effects (EPA, 2005). There are a set of standard statistical methods for comparing proportions, such as tumour incidence, between one or more groups. In these tests the basic information is the number of animals at risk as the denominator and the number of animals with the tumour (or pathology) as the numerator of interest. These tests make no assumptions or corrections to take into account 'COO' or 'COD' information or the time to tumour or death. They are described in many standard statistical textbooks and software for analysis is widely available.

341. The three main tests are:

1) A pair-wise test between the negative control and the treated group using the Fisher exact test (Fisher, 1950).

2) A chi-square test for heterogeneity of proportions between groups. This is an 'omnibus' test of differences between a series of groups (with no ordering to the groups).

3) A test for a linear trend such as the Cochran-Armitage trend test (Snedecor & Cochran, 1980).

342. Pair wise comparisons are carried out using Fisher's exact test or its Chi-square approximation. Fisher's exact test is now preferred because of the availability of software for carrying it out.

343. The Cochran-Armitage trend test aims to detect a linear trend. The null hypothesis is that all animals are at equal risk of developing a tumour during the study. Problems arise if there are differences in mortality between the groups. The test is sensitive to increases in treatment related lethality and this leads to an incorrect level of the Type 1 error (the risk of falsely rejecting the null hypothesis).

344. The selection of the dose metric (the values representing the actual doses used) for use in the analysis is important. The doses could be the 'applied', logarithmic, some equally spaced rank or some measure of the effective dose at the target organ (based upon pharmacokinetics). A practical issue arises using a logarithmic scale with the need to choose a value to substitute for the zero dose level.

345. A trend test is more powerful than the pair-wise test. A complication is that a trend test may fail to detect curvi-linear responses such as might arise from non-linear effects such as complications from saturation. In such situations the pair-wise tests may give more appropriate results.

4.16 Survival adjusted analyses

346. For the reasons discussed above, survival adjusted methods are strongly advocated for comparisons of tumour incidences among groups.

347. Some exact probability test methods are available for use instead of the approximations based upon the normal distribution used in the standard methods. The standard methods may underestimate p values when the numbers of tumours in the groups are small. In these situations exact permutation tests which are extensions of the Fisher exact test are suggested for use (Lin, 2010).

348. Three different types of statistical procedures have been developed depending upon the type of tumour.

1) The prevalence method for non-lethal (incidental) tumours (Hoel & Walburg, 1972; Peto *et al.*, 1980)
2) The death rate method for lethal (fatal) tumours (Tarone, 1975; Peto *et al*., 1980)
3) The onset–rate for mortality independent (observable) tumours (Peto *et al*., 1980)

The prevalence method

349. The prevalence method, also called the incidental method, is, effectively the Hoel-Walburg procedure for nonlethal tumours which makes no assumption about the COO of the tumour (Hoel & Walburg, 1972).

350. The procedure involves carrying out a life-table analysis, a method for analysing censored observations that have been grouped into intervals, for tumours which were found incidental (incidental context) to the death of the animal. The experimental period is split into a set of intervals (including any interim or terminal sacrifices).

351. It has been argued that the choice of partitions is not critical. Peto *et al*., (1980) suggest the partitions should not be so short that the prevalence of incidental tumours is unstable nor so large that the prevalence could differ markedly from one half of an interval to the other and suggest a partition based upon 'ad hoc' runs. However, there have been concerns about constructions of ad hoc intervals in the Peto analysis and attempts made to standardize them. The US National Toxicology Program (NTP) for instance previously used intervals 0 – 52, 53 – 78, 79 – 92, 93 – 104 plus the terminal kills.

352. In the analysis the denominator is the number of animals dying within the specific partition and the numerator is the number of animals dying with an incidental tumour. The analyses for each individual partition are combined using the Mantel-Haenszel method. This compares two groups for proportions with an adjustment for control variables. A series of k 2 x 2 contingency tables are produced with k being the number of strata of the different control variables such as age or sex. The stratification increases the power of the design to detect an association. The test statistic is a chi-square. Full details of the equations used to conduct it are found in Lin (2010).

353. The methodology uses normal approximations for the tests but the approximation may be unreliable if the number of tumours in a group is small. A permutation test involving the hypergeometric distribution may be used (Lin, 2010).

The death rate method (for comparing rapidly fatal tumours)

354. This test is also referred to as the log rank test assuming that all the tumours cause death (Peto, 1974). It is used when tumours are observed in a fatal context. In the analysis the stratification is based upon a partition into intervals (often a week) where one or more animals died. In each

stratum the number of animals entering the partition or strata who are tumour free is the denominator and the number of animals dying during the partition with a fatal tumour is the numerator. The analyses for each individual partition are combined using the Mantel-Haenszel methods. Full details of the method and equations are found in Lin (2010).

Mortality independent analysis (onset rate method)

355. In those tissues such as skin and the mammary gland where tumours can be observed in the live animal, the tumours are described as mortality independent. The onset rate method (the same basic statistical method as used for the death rate or fatal analysis) is used for these mortality independent tumours. The endpoint in this case is the occurrence of a tumour based upon it reaching some predefined size rather than the death of the animal. Once a tumour has been identified using this criterion, the animal makes no further contribution to the analysis. It is no longer 'at risk' of developing a tumour because it now has a tumour even though it may live on for some time. The calculations used to produce the statistics for the onset rate methods are comparable to those produced by the death rate method.

Peto / IARC analysis

356. The Peto analysis, as described in an IARC monograph (Peto *et al.*, 1980), is a combination of the prevalence and death rate based upon the COO of a tumour as either being called non-lethal ('incidental') or the cause of the animal's death ('fatal'). It is a joint test of age-adjusted tumour lethality and age-adjusted tumour prevalence. The assumptions underlying the method are that the control and treated animals are equally likely to be killed at any particular stage in a tumour's development, the animals dying of other causes are representative of all animals surviving in that interval and that the pathologist is prepared to make the 'fatal'/'incidental' classification.

357. The analysis combines the two separate approaches. Animals whose tumours are termed 'fatal' are analysed by Peto's death rate method while those called 'incidental' are analysed using the prevalence method. The two analyses are then combined using the Mantel-Haenszel method to provide a test for trend. A problem is that the analysis will be biased if the assumption on the nature of the tumour being either called 'fatal' or 'incidental' is inaccurate (Dinse, 1994). Although the Peto test is described as robust, there is some debate as to how big a problem misclassification is.

358. Peto *et al*. (1980) provide an illustration of the implications of wrongly defining the COO. For example, in a study of pituitary tumours in animals treated with N–nitrosodimethylamine (NDMA) different conclusions are drawn depending upon the COO. If all the tumours were considered fatal then NDMA was wrongly considered carcinogenic, whereas the tumours were considered incidental then NDMA was wrongly considered health protective. Taking the COO into account produced what Peto *et al.* considered to be the correct interpretation that there was no carcinogenic effect in the pituitary.

Multivariate regression methods

359. Statistical methods used for survival analysis such as the Cox proportional hazard method have been used for the analysis of carcinogenesis data. This is a regression method which takes into account the time until the binary event of interest (death) occurs. The modelling also allows the effect of an exposure/treatment to be investigated after adjusting for confounding effects. The method was applied when the tumours were considered lethal.

360. Dinse and Haseman (1986) suggested a logistic regression approach which was applicable when the tumours were considered incidental, Logistic regression is a method used where the

outcome is binary such as whether an animal has a tumour or not. The effect of an exposure on this binary outcome can be adjusted for confounding factors in a study.

Poly-k test

361. More recent approaches have been the development of the poly-k tests (Bailer & Portier, 1988; Dinse, 1994). The poly-k test does not need arbitrary partitions of time periods or COO information. The test is based upon the assumption that the time to tumour onset can be modelled based upon the tumour onset times raised to the power k. Initially, the test was proposed without identifying how to derive k but now it is suggested that k should be 3 because of observations that tumours can be modelled by a polynomial of order 3 from an analysis of NTP historical control data for F344 rat and B6C3F1 mice (Portier *et al*., 1986.) The poly-3 test is then a special case of the poly-k test. The power of 6 (or k=6) can be used when the tumour onset times are close to a polynomial of order 6. The value of k need not be critical as poly-k tests are reported to give valid results if the true value lies between 1 and 5 (Bailer and Portier, 1988).

362. The tests are, in effect, modified Cochran-Armitage tests which adjust for differences in mortality in the treated groups by a modification of the number of animals in the denominator to reflect the less than whole animal contributions because of reduced survival. The approach gives a value (w) from 0 to 1 for each animal based upon a weight which relates to the time of death or the time of the final sacrifice. Thus w relates to the fraction of the length of time the animal survived in the study over the total length of the study to the power k. The value w is <1 if the animal died early without developing a tumour and w = 1 if the animal died with a tumour or survived until the study was completed. The number of animals at risk is replaced by a new estimate in the Cochran Armitage test. The method tests for a dose-related trend in the mortality-adjusted lifetime tumour incidence rate.

363. The Bieler-Williams variance (1993) is used in the poly-3 test (which is also sometimes referred to the Bieler-Williams method/test) where the test is modified by using the delta method and weighted least squares techniques in order to adjust the variance estimation of the test statistic and to improve the performance of the test.

Comparison between Peto and Poly-k methods

364. Rahman & Lin (2008) compare the false positive rates of the Peto and poly-k tests using a simulation study. Kodell, as cited in Lee *et al.* (2002), compared the properties of the Peto and poly-k tests and concluded that both are valid for adjusting for differential mortality. The problem remains that the comparison is ideally based on tumour onset data but because most tumours are occult, the tumour onset cannot be actually observed. In addition, there is a need to adjust the analysis for any differences in inter-current mortality.

365. The NTP routinely used to carry out two trend tests. One assumed that all tumours in dead or moribund animals were 'fatal'; the other assumed all the tumours were non-fatal ('incidental'). The current NTP approach is to no longer use life-table tests or prevalence tests. Instead, the poly-3 test with Bieler-Williams variance with a trend test and pair-wise tests with controls is used. Sometimes this test is used with k=1.5 and/or k=6.

366. Debate over the methods for the analysis of data on the COO of tumours using the Peto analysis has generated controversy. The Society of Toxicologic Pathology (STP) set up the STP Peto Analysis Working Group and produced draft recommendations for the classification of rodent tumours for the Peto analysis. Instead of the Peto analysis the STP Working Group recommended that the poly-3 methodology be used (STP 2001).

367. STP concluded:

1) Pathologists cannot determine the time of onset accurately from post mortems.

2) If the Peto analysis uses death as a surrogate for time of onset then the method seems inappropriate

3) Better to use other methods which do not require COO information

368. Lee & Fry (cited in Lee *et al.*, 2002) responded to the STP Peto Analysis Working Group recommendations and their comments together with those of other statisticians were published as a collection of comments in Toxicologic Pathology (cited in Lee *et al.*, 2002).

369. Kodell (cited in Lee *et al*., 2002) concluded that both the Peto and poly-3 tests are valid for adjusting for differential mortality. Both methods are fairly robust to deviations from their assumptions although both could be improved by modifications and further development.

370. Based upon these comments particularly relating to issues around the use of the Peto method and the onset of fatal tumours, the STP withdrew its criticism of the Peto approach while maintaining that the poly-3 test is appropriate in certain circumstances (STP, 2002).

371. The STP (2002) put forward new recommendations that:

- The Peto test should be performed whenever study pathologists and peer review pathologists can consistently classify neoplasms as fatal or incidental

- If fatal and incidental classifications are not applied, the Poly-3 or another alternative to the Peto test should be employed

4.17 Tests of difference in survival

372. As discussed earlier (see paragraph 326), differences in survival can affect the conclusions drawn from the simple analyses. A number of methods can be used to test for differences in survival and for significant dose-response relationships or trends in studies. These include: the Cox test (Cox 1972; Thomas, Breslow, and Gart, 1977; Gart *et al*., 1986); the generalized Wilcoxon or Kruskal-Wallis test (Breslow 1970; Gehan 1965; Thomas, Breslow, and Gart 1977); and the Tarone trend tests (Cox 1959; Peto *et al*., 1980; Tarone 1975).

373. The number of animals surviving until the scheduled terminal kill can be compared by what is termed the product-limit or Kaplan-Meier method (Kaplan and Meier, 1958) and plotted graphically. This analysis compares length of survival in the groups (and does not involve any pathology findings.) Animals that are found dead other than from natural causes are usually treated as censored and excluded from the survival analysis. There is a pair-wise comparison method (Cox, 1972; Tarone & Ware, 1977) and a test for linear trend, Tarone's life-table test (Tarone, 1975).

374. The Kaplan-Meier analysis involves calculating the ratios of the surviving animals divided by the numbers of animals at risk of dying. Every time an animal dies the ratio is recalculated. These ratios can be plotted to show a curve which displays the probability of survival. When there are different treatment groups a curve can be generated for each group. Formal statistical tests such as the log rank test can be used to test difference between groups. (Curves which are close together may indicate that a difference is not statistically significant.)

4.18 Assumptions for statistical analysis

375. In the context of the theme of this document, it should be reiterated that the design of the experiment is fundamental to the choice of statistical methods.

376. Statisticians have long had a specific interest in the methods used to allocate animals to treatment, sample size determination, the control of possible confounding effects such as cage location and their rotation, the dose selection and the length of the study (Haseman, 1984).

377. An important consideration is the need to balance the considerable experience built up in the use and interpretation of the data from these experimental studies over many years with the appreciable empirical knowledge that exists on the implication of violating the assumptions associated with the statistical methods.

Randomisation

378. One assumption underlying the design and subsequent statistical analysis such as the ANOVA is that the animals have been assigned at random to the treatment groups. Each animal entering the study should have the same chance as any other of being allocated to one of the experimental groups (including the interim kills and control groups). Randomisation can be carried out by a number of different methods. The primary objectives are to prevent bias and to ensure that uncontrolled covariates do not affect the results of the analysis.

379. Stratified randomisation with groupings based upon body weights may be used to reduce bias and ensure compatibility of the various treatment groups with respect to uncontrolled variables. Ideally the animals from all groups should be placed into the study at the same time. If this is not practical, strata or blocks can be included as factors in the subsequent statistical analyses and can be created by starting subgroups from the control and each of the treatment groups over several days.

380. An assumption in the statistical analysis is that randomisation occurs at all points in a study so that biases are not introduced. If, for instance, the animals are initially randomised into groups but subsequent procedures are carried out in a pre-defined systematic order there is a risk that biases may be introduced.

Independent experimental units

381. The animal is often both the experimental and observational unit. If a cage is assigned to a treatment it becomes the experimental unit. However, the common statistical methods used in the analyses make the assumption that the experimental units are independent. In some cases this is clearly not so. In practice, although the assumption of independence of experimental units is an important one (more important than normality and homogeneity of variances; van Belle, 2009) it is not always taken fully into account in toxicological studies.

382. Individual housing is preferable to meet statistical assumptions but may have implication for the welfare and the representativeness of the animal. The trade off is between a theoretical optimum design, the avoidance of cage/confounding effects and practical husbandry issues, the 'pathology' of single housing and the possible increased variability of isolated animals. Litter and caging effects should be taken into account in the statistical analysis. If this cannot be done this should be noted with an explanation of why doing this is not possible, and the potential implications for the study.

383. Lack of independence may also arise from contamination of doses within and between cages. In the case of airborne contamination the random assignment of cages will mean that some of the lower dose and control groups will be exposed to some or higher concentrations than is implicit in

the experimental design. There is clearly a need to balance the potential limitations resulting from contaminations arising from randomisation in a room (but which also provides some protection from this uncontrolled variable) with the potential confounding effect of separate handling and housing of dosed groups necessary to prevent cross treatment group contamination.

4.19 One- or two-sided tests

384. The choice of whether to use a one- or two-side test should be made at the design rather than the analysis stage. A two-sided statistical hypothesis test tests for a difference from the negative control (in a pair-wise comparison) in either direction. A one-sided comparison tests for a difference in only one pre-specified direction, but as a consequence has more power. In a carcinogenicity study, the expectation is often that the change will be an increase in tumours in the treated group so a one-sided test may be considered more appropriate, although this can be controversial. If the treatment could also be protective (i.e., reduce tumour incidence or delay it) then a two-sided comparison may be more appropriate. Regulatory authorities may have specific opinions. For instance, the US EPA (2005) notes that either "a two-tailed test or a one-tailed test may be used".

Equal information per unit

385. There are implications for the assumptions underlying the statistical analysis in the non-random reading of histopathology slides. In some studies more effort has been put into reading slides from the control and top dose groups with less emphasis on the examinations at the intermediate doses. This approach was more common when qualitative hazard identification was the objective. Although such studies have been accepted by some regulatory authorities, this method can create problems in the statistical analysis of dose-response trends and cannot be recommended if dose-response characterization is an objective of a study. Similar considerations arise if, because of uncertainty in a diagnosis, more slides are read for some animals than others.

Blind or unblind reading of slides

386. Blinding procedures such as for the reading of slides is discussed in paragraph 258 (section 3.6.5). Many statisticians expect blinding to treatment group to be included, as in clinical trials as a protection against biases. Temple *et al*. (1988) have argued that blinding should be used to guard against biases that might arise (inadvertently) from the pathologist knowing the treatment group of the animal from which the tissue was derived. However, Haseman, in the STP discussions (cited in Lee *et al.*, 2002) argues from a statistical point of view, that blinding is not necessary given the long experience of using the assay. Blinding is, however, often used when re-reading slides in disputed cases or when the results are close.

Confounding variables

387. One of the purposes of randomised and blinded studies is to minimize the effects of uncontrolled covariates and to prevent the introduction of biases by preventing confounding factors from distorting the results. A confounding variable is one which is so closely related to both treatments (or another factor in the study design) and effects such that the individual contributions to the effect of interest cannot be separated. Haseman (1984) discusses a number of confounding effects such as cage location and litter mates although the latter should be controlled by randomisation. Butterworth *et al.* (2004) discuss another form of confounding where, rather than the chemical under study, a contaminant may be responsible for the carcinogenicity detected.

388. In the case of long-term bioassays, early differences in body weight between control and treated animals which persist through the study create a potential confounding effect between body

weight, life span and tumour incidence. Ad libitum overfeeding is the most important uncontrolled variable which affects the results of a long-term bioassay. Keenan *et al*. (1996) report, for instance, that there is a highly significant correlation between food consumption, body weight / obesity and shortened life-span in rodents. Kodell *et al*. (2000) notes that the reduced survival of Sprague Dawley rats questions the continued use of this stock in the carcinogenicity study.

389. If the tumour profile of lean and obese rodents is different, then there is a possibility that apparent treatment related differences in tumour incidence may, in fact, be wholly or partially caused by the body weight differences. Confounding may then make it difficult to identify if an effect was a direct result of the treatment or an indirect effect of the treatment through affecting food intake and consequently body weight. Ibrahim *et al.* (1998) noted that there is usually not enough information to provide a regression based adjustment for differences in body weight.

4.20 Interpretation of statistical analyses

390. Interpreting the results of a long-term bioassay is complex. A critical issue is the practical problem of the low power of the design when the tumour incidence is rare together with the multiple comparisons issues arising from the investigation of 20 or more tissues from both sexes of two species. As a result, there is a risk of both Type I (false positives) and Type II (false negative) errors. It is also necessary to integrate the results of the full battery of statistical tests, significant or otherwise, and the importance of a series of biological issues in the assessment of the result. Factors which add importance to findings include: uncommon tumours, multiple sites, positive findings using more than one route of administration, effects in multiple species/ strains/ sexes, effects at the same site in both sexes and/or species, tumour/disease progression, increased preneoplastic lesions, reduced latency, metastases, unusually large responses, dose-related responses and a high proportion of malignant tumours. In addition, a comparison of the observed tumour incidence in the study with the historical control rates should be made.

391. Various approaches are taken to control the false positive rate. Haseman (1983) developed a decision rule. This approach is based upon comparisons between the top dose and the negative control using data from the standard NTP experimental design studies using F344 rats and B6C3F1 mice. The rule uses a criterion of a statistically significant difference at $P<0.01$ for common and <0.05 for rare tumours (Lin & Ali, 1994). The definition of a rare tumour is an incidence of <1%, based on historical controls. Above 1%, the tumours are considered common. It has been argued that Haseman's rule produces too high a false positive rate because all treatment groups rather than just the top dose are, in fact, used in the comparisons. Any decision rule used to control the false positive rate should be justified.

392. A study where no treatment–related effects are found should be reviewed to ensure it is valid by checking that sufficient numbers of animals lived long enough to ensure that adequate exposure had been achieved and so were 'at risk' of developing 'late in life' tumours. A check should also be made that the objectives of the study design in terms of dosage have been achieved (see section 3.1 on dose selection).

4.21 Use of control data and dual control groups

393. OECD TGs 451/452/453 specify that a concurrent control group should be included in the study design. Study designs where there are more than one group of concurrent controls are rarely necessary, but when used, come in two basic types: firstly, where the two control groups are treated differently, e.g., vehicle contol and an untreated control group or a pair fed control vs ad libitum (see paragraph 283); secondly, where there are two identical control groups (Haseman *et al*., 1986).

394. With the first type, the objective is to see whether differences between the 2 control groups have any effects on tumour incidence or any influence on any other toxicological effects in the control animals. These different control groups should not be combined for use in statistical analyses. The concurrent control group that only differs from the test groups by the absence of the test substance is the most appropriate for the comparison with the test groups.

395. The argument made for the second type, the identical dual controls (C1 and C2) is that they provide a way of identifying the degree of variability in the negative control animals. This gives a better basis for addressing the biological importance of any increase found in the treated groups. They can be considered as 'contemporary' historical control data.

396. Data from the two dual control groups (C1 and C2) can usually be combined (Haseman *et al.*, 1990) but if differences are found either in mortality or tumour incidence then it has been suggested that three tests with the treatment group (T) could be carried out (C1 v. T) (C2 v. T) and ((C1+C2) v T).

397. Some have argued that if the C1 and C2 groups differ then comparisons with the test groups are only considered positive when comparisons with each control group individually is significant on the grounds that findings should be reproducible. However, if it is accepted the carcinogenicity study is underpowered, then the comparison would be considered positive if any of the comparisons of the treated group with the control groups were significant. The first approach risks more false negatives, the second approach more false positives.

4.22 Historical control considerations

398. Historical control data can help interpret results in a number of situations (see GD 35). In any discussion about historical control data, it should be stressed that the concurrent control group is always the most important consideration in the testing for increased tumour rates. The historical control data can, though, be useful provided that the data chosen are from studies that are comparable with the study being investigated. It is widely recognized that large differences can result from disparities in factors such as pathology nomenclature, strain, husbandry, pathologists.

399. It has been suggested that historical control data should only be used if the concurrent control data are appreciably 'out of line' with recent previous studies and that only historical data collected over the last 5 years should be used. Such historical control data can be helpful in evaluating how 'normal' the data from the concurrent control groups are, for evaluating differing results from the dual control groups and as a form of quality control for carcinogenicity studies. Any concerns over the appropriateness of the control groups need to be evaluated and discussed.

400. Elmore & Peddada (2009) discuss how to incorporate historical control data into the statistical analysis of the carcinogenicity study. The mean and SD can be affected by a 'rogue' outlier while the median and interquartile range (IQR) is not. They argue outliers should be identified and not discounted but considered alongside other relevant data in the assessment of the results. They suggest the use of exploratory methods such as box and whisker plots (with their associated 5 number summaries) to give graphical presentations of historical control data (when there are more than 15 studies) together with the results of the treated groups and the concurrent negative control in a study. They illustrate the advantages of using the median and quartiles over the mean, SD and range especially to identify potentially misleading outlying results. They recommend using Bailer & Portier (1988) survival adjusted tumour incidence rates.

401. There are a number of statistical methods developed for incorporating historical control data into formal statistical analysis. Although this is an area of active statistical research, these methods have not been used in a regulatory context.

402. A number of methods have been described for incorporating historical control data into the statistical analysis of a trend in the data. One suggestion has been to use the upper confidence limit on the binomial proportion to help in interpretation. Tarone (1982) developed a method using the beta-binomial distribution (a binomial distribution where the value of p is a random variable rather than a single fixed value) to account for the variability between studies to model historical control data and to derive both exact and approximate tests. Including the historical control data increases the power of tests, especially for comparisons with rare tumours but the method does not take into account differential survival. Ibrahim & Ryan (1996) have developed a test which uses historical control data in survival-adjusted tests. The study period is split into intervals and in each of these the multinomial distribution is used to model the number of animals dying with tumours. The prior distribution for the historical control rate is based upon the Dirichelet distributions. The method, though, can only be applied to fatal tumours. Ibrahim *et al*. (1998) developed methods for incorporating historical control data into age-adjusted tests. This approach, though, is limited because it makes strong assumptions about the tumour lethality.

403. Ibrahim and co-workers have also developed a method that assumes all the tumours are lethal (Ibrahim & Ryan, 1996) or all tumours are non-lethal (Ibrahim, Ryan and Chen 1998). The two tests, therefore, represent the extreme events and may not be accurate in practice. Fung *et al.* (1996) developed methods for incorporating historical control data but this approach is also not an age-adjusted test.

404. A Bayesian approach has been suggested by Dempster *et al*. (1983) making the assumption that the logits of the historical control rate are normally distributed. Another Bayesian approach has been developed by Dunson & Dinse (2001) which relaxes some of the assumptions regarding the nature of the tumours. The prior probabilities for the parameter in the model, however, have to be chosen carefully and this requires a consensus between the pathologist and the toxicologist.

405. Peddada *et al.* (2007) have proposed a non-parametric approach which provides separate comparisons of the dose groups with the concurrent and historical control data. A third comparison can be made between the two control data sets. The poly-3 correction is made to sample sizes to adjust for survival rate differences between the groups. Consequently, individual data rather than summary data for a group are needed. The three p-values obtained are compared using a 'weight of evidence' approach. Without the survival data there is a possibility of bias.

4.23 Dose-response modelling

406. Previously, most of the emphasis on the carcinogenicity study had been on hypothesis testing and the identification of a carcinogenic hazard. The tests have been about the identification of a significant effect or trend. An alternative approach is the estimation of the size of any dose-response relationship. However, the evolution of the carcinogenicity study towards greater emphasis on quantitative dose response modelling has implication for the study design. The traditional design is limited because of some of the other objectives described earlier (paragraph 272).

407. The current approach to experimental design is a trade-off between the optimum allocation of equal group sizes between a control and treated group to maximize the power to detect an effect (i.e., test a null hypotheses) and the need to describe in detail the shape of a dose response relationship. Studies need to be designed to identify which parts of the relationship are important and, consequently, may require a two-stage approach of identifying a dose range of interest in a

preliminary range finding study and then moving to investigate this in more detail by concentrating resources there.

4.24 Extrapolation to low doses

408. In the current paradigm of risk assessment, the objective is to identify the risks associated with levels of potential human exposure. Risk assessment has traditionally been carried out differently for genotoxic and non-genotoxic carcinogens and non-carcinogens. A no-threshold model has usually been assumed for genotoxic carcinogens which has been associated with low dose extrapolation. On the other hand, a threshold model has been assumed for some non-genotoxic carcinogens and non-carcinogenic endpoints, provided there are adequate supporting mode of action data. The threshold model, below which there are no observed toxic effects, has been linked to the concept of a point of departure (POD) e.g., a no observed (adverse) effect level (NO(A)EL). In the absence of sufficient mode of action data, the risk assessment approach usually reverts to the non-threshold model.

409. Low dose extrapolation has been carried out by some regulatory agencies by fitting mathematical models to the observational data for carcinogens and then extrapolating the models to the low doses/exposures that might be expected to occur in the human population (EPA, 1986). For a long time, The US EPA used the Linearized Multistage procedure as the default approach for such extrapolations. These approaches aimed to identify, for instance, Virtually Safe Doses (VSDs) where it has been estimated that such life time exposure would lead to an upper bound increase of 1 extra lifetime cancer death in 1 million exposed individuals (or a 10-6 lifetime risk) or some similar low risk. The low dose extrapolations are, however, highly dependent upon the mathematical function assumed for the dose-response relationship and could give very different estimates of, for instance, the VSDs. Such low dose extrapolation may be conservative and alternative approaches have been proposed. Modelling may be confined to the observed experimental ranges and a POD identified such as, for example, an estimate of the Benchmark Dose.

410. Some authorities now propose linear extrapolation from the POD by drawing a straight line from the POD to the zero extra/additional risk and reading off the VSD associated with the 10^{-6} excess/additional risk (US EPA, 2005). Others propose using the ratio of the POD to the human exposure to derive a Margin of Exposures (MOE).

4.25 NOEL, NOAEL, LOEL, LOAEL approach

411. A NOEL (no observed effect level) is obtained by identification of the highest dose level of a test substance that does not cause a significant increase in any treatment-related effects compared with the negative/vehicle control. The LOEL (lowest observed effect level) is the lowest dose where there is a statistically significant effect.

412. The NOAEL (the no observed adverse effect level, sometimes designated NO(A)EL) is the highest level of a test substance that does not cause any observed and statistically significant adverse effects compared with the controls. The NOAEL distinguishes between changes which are adverse rather than any treatment related effect which may in some case not be adverse. Similarly the LOAEL (sometimes referred to as LO(A)EL) is the lowest dose where there is a significant adverse effect compared with the controls. Other terms used are NOEC, NOAEC, LOEC and LOAEC where the C refers to concentration rather than level, e.g., in inhalation studies.

413. The NOAEL identified in a long-term toxicity or carcinogenicity study may be used to derive a health-based guidance value such as an Acceptable or Tolerable Daily Intake (ADI and TDI), by applying a Safety or Uncertainty Factor (SF or UF) to the NOAEL derived from the study. Application of Safety or Uncertainty Factors depends of regulatory context. It is then considered that

there are no appreciable health risks below this health-based guidance value (WHO, 1999). The SF/UFs are used to account for inter- and intra-species variability, and an extra SF is applied if no NOAEL can be found and the LOAEL is used to derive health-based guidance values. The NOAEL approach is only relevant in the case of those non-genotoxic carcinogens or non-cancer endpoints where there is believed to be a threshold dose below which no toxic effects occur.

414. Although widely used, the following limitations of the NOAEL approach have been suggested:

- The NOAEL is inherently dependent on the study design, sample sizes, dose spacing, background levels, the power of the design (and statistical test)
- It is also based upon a hypothesis testing approach where a failure to reject the null hypothesis is taken as evidence of no difference. This differs from how a failure to reject a null hypothesis is interpreted by statisticians,
- The statistical comparisons are pair-wise comparisons and do not take into account information on the dose-response relationship or the more powerful trend test.
- The smaller the experiment or the more variable the endpoint the less chance of detecting a real effect
- The NOAELs tend to be higher for endpoints with a high control/background level because it is more difficult to demonstrate a statistically significant difference than for endpoints with low control/background levels
- It is a single dose level and does not reflect the other dose-response information e.g., the shape of the dose-response curve
- The NOAEL is not a measurement as such so cannot be presented with a measure of precision such as a confidence interval (CI).
- The NOAEL is a dose at which no adverse effect was detected. This does not preclude effects that could not be detected by the techniques used and also does not indicate that higher doses below the LOAEL may also have no detectable adverse effects. It is a single dose level dependent on the study design and is not necessarily representative of the true 'threshold'

4.26 Benchmark dose approach

415. An alternative approach to the NOAEL approach, the Benchmark dose (BMD) approach, was first proposed by Crump (1984) as an alternative for the identification of estimates of dose levels helpful for risk assessment. It had a slow acceptance despite early work to generalize the concept but has gradually become more widely accepted.

416. The BMD approach is based upon fitting a mathematical model to dose-response data collected for either qualitative or quantitative endpoints from experiments with 3 or more dose groups plus a negative control group. Mathematical models are fitted to the observed data. The dose associated with some predefined increase over background is chosen as the BMD. The lower confidence limit on the dose is then chosen as the POD for further investigation of the likely response at lower dose depending upon the choice of method to be used for extrapolation to lower doses.

417. The modelling approach involves the choice of endpoints to model, the choice of an appropriate mathematical model to fit to the data, the selection of model(s) which are considered to provide a satisfactory fit to the observed data and the selection of an appropriate BMD for determination of the POD.

418. The BMD is the dose associated with a pre-specified change in response. This response, the BMR, is a change in the endpoint of interest above the control response. In its original form, the response was based upon an increase over the control incidence of a quantal measure (such as tumour incidence). Values such as a 5% or 10% additional or extra incidence have conventionally been used to define the BMD for quantal data.

419. The use of confidence limits in the BMDL approach provides an estimate of uncertainty, that the uncertainty is reduced in large studies with better designs usually leading to higher PODs. It uses all the data in a dose-response experiment. The POD does not have to be one of the experimental dose levels and a POD can be calculated even if there is no NOAEL derived from a study.

420. Acceptance of the BMD approach is not universal. Travis *et al*. (2005), for example, argue against its routine use and that there are issues about how to apply it in cases where there is either no LOAEL identified (i.e., no obvious dose-response) or no NOAEL is identified (significant effect at all dose levels). Travis *et al*. (2005) argue that the NOAEL is best for routine use in toxicology studies but that the BMD may have a role in the interpretation of the most influential/critical studies in a regulatory package. The BMD approach also has potential limitations in that the selection of the model type and the parameters to include are chosen by the 'assessor'. This means that the same data may produce different BMD/BMDLs. There may also be difficulty in arriving at a consensus in defining a BMR level of a measure. The BMD, just as the NOEL, is also not a risk- or response-free exposure level.

Mathematical modelling for the BMD

421. Those endpoints which show 'visual' trends are analysed further to identify if the dose-response data are suitable for further analysis by fitting dose-response models to the experimental data. After suitable models have been identified, the BMD and the BMDL are determined for each suitable model. The specific BMD value is determined either by interpolation within the experimental data or by extrapolation beyond the experimental data. The BMDL is the lower one-sided 95% confidence limit on the BMD value. The BMDL can be interpreted as meaning that there is 95% confidence that the true effect at this dose would be less than the effect associated with the BMDL.

422. Values are obtained using software (BMDS developed by the US EPA (http://www.epa.gov/ncea/bmds/index.html) or the S-plus and R based PROAST software developed by RIVM) specifically developed for the purposes or using statistical packages such as SAS. (It is important that the various assumptions and defaults underlying these approaches and, in some cases, incorporated in the software are acknowledged.) One assumption in the modelling approach is the distribution of the data. In the case of quantal data the assumption is a binomial distribution. In the case of quantitative data, either a normal or a lognormal distribution is assumed.

423. The modelling approaches use algorithms to identify the optimal values of the parameters which specify the mathematical model. These values are derived by minimizing the difference between the fitted values and the observed values. One approach is the maximum likelihood method where likelihood is a measure of how likely the parameters have these specific values given the observed data; parameter values that maximize the likelihood measures are considered the 'best' estimates'.

424. The BMDL is usually estimated using a likelihood ratio test, a method also used for comparing the fits of different models. Twice the absolute difference between the log likelihoods of two different (comparable) models follows a chi-square distribution with degrees of freedom equal to the difference in number of parameters in each model. A chi-square value significant at $P<0.05$ is taken by some as evidence that the two models are considered significantly different.

425. The range of models potentially suitable for quantal data includes the 'standard' tolerance functions: probit, logistic and Weibull and log based versions. Others include the multistage and gamma multi-hit models. More complex models are provided in the software to take into account the more complex multilevel data from developmental/teratology data where there are, for instance, intra-litter correlations.

426. A large number of models can be used to describe a dose-response relationship in the case of quantitative (continuous) data. Examples include the polynomial, power, exponential and Hill function models (Slob, 2002). The Hill and exponential models have the properties of being sigmoid (S-shaped) and bounded (levelling of at a particular maximum and minimum response value) and parameters that can be easily related to the shape of the dose-response relationship. (On a log scale the Hill model is symmetrical but not on the normal dose scale.). Models can also be specified to take into account different variances within the dose groups.

427. A sufficient number of dose groups (with different response levels) are needed to assess whether a dose-response is linear, sublinear, sigmoidal or another form. In the case of a sigmoid/S-shaped dose-response, the modelling requires at least five dose levels (including controls) to avoid over-parameterization (perfect fit) and allows a test of model fit. A family of a specific model, e.g., polynomial, linear, power, Hill and exponential, is often used in the BMD approach for continuous data. . Testing within the families can be carried out with a likelihood ratio test for whether extra parameters improve the fit. If they do not, then by parsimony, they are left out. If doses were not high enough to measure where a sigmoid dose-response relationship levels off, the approach of nested models would fit a simpler model, without the need of identifying the maximum response level. Including high doses (where the response levels off) would result in more precise BMD estimates, but they are not crucial.

428. There is complexity in comparing models across classes. The US EPA has suggested the use of Akaike's Information Criteria (AIC). The AIC is a measure of the fit of the model weighted by the number of parameters fitted with the model, with the lowest AIC selected. A complication is that, when models are similar, the relative ranking in terms of AIC may be somewhat arbitrary. In practice, however, the objective is not to find the "best model" but rather to identify models which are plausible.

429. In practice, different models with the same number of parameters can often be found to give a satisfactory fit to the same dataset. The approach then is to calculate the BMDs and BMDLs from these various mathematical models and compare the range of values for all acceptable models for their similarity and consistency. The choice of which model is used for subsequent calculations may be based upon criteria such as which approach gives the lowest or most conservative BMDL or gives the best visual fit to data. Such choices, therefore, have potentially appreciable input from the risk assessor.

430. For the selection of the BMDL, the case is made for the selection of the lowest (i.e., most conservative) BMDL from those models which fit the data satisfactorily. Model averaging has also been proposed using, for instance, Bayesian model averaging where the averaging is a weighting derived from the support for a particular model taking into account the data (Bailer *et al*., 2005). The BMDL which is considered most appropriate is identified as the RP which is used by the different approaches to defining a guidance level for risk assessment (ADI, TDI, MOE etc).

431. In the case of quantal data the BMR may be expressed either as 'additional' or 'extra' risk. Extra risk is an adjusted rate which includes an adjustment for the background incidence rate and is based only on the fraction who are expected not to have a background incidence with BMR = P(BMD) – P(0) / 1 - P(0). Additional risk is an absolute rate: P(BMD) – P(0). The two terms become the same when P(0), the background frequency, is zero.

432. The extra or additional risk (BMR) used to derive default BMDL values has usually been 5 or 10% for quantal data based upon the similarity between the BMDL derived from them and the NOAEL derived from developmental toxicity studies. The BMDL for a 10% risk level initially seemed most similar to the NOAEL derived from the same studies. Using more complex models taking into account the intra-litter correlations suggested the BMDL for a 5% risk level were most similar to the NOAEL.

433. One limitation of modelling for carcinogenicity study data is that it is generally done solely on incidence (quantal) data and does not make use of time-to-tumour (continuous) data. Including these data should be both more informative and provide a more accurate assessment. However, modelling methods using them have not been developed and validated and this limitation is particularly important when there is uncertainty over the cause of an animal's death. Other complexities for such methods are the consequence of early termination of studies or of some groups.

434. In the case of quantitative data the percentage change of effects relative to control levels that are considered biologically important need to be defined. There have been a number of different definitions of the BMR based upon different underlying approaches (IPCS, 2009)

435. In one approach the BMR is related to the CES (critical effect size) (Slob & Pieters, 1988). The CES is a BMR defined as a percentage change in the average response compared with the average response level in the controls: e.g., a 10% change, in the mean body weight over the control values for an adult animal or some fold change in an enzyme level or clinical chemistry value. There remains a debate over what the CES should be for each toxicological endpoint and the potential to use within animal variability to define it (Dekkers *et al*., 2001; 2006). A 5% CES has been suggested in the absence of other information (Woutersen *et al*., 2001; EFSA, 2009) in part because such a CES seemed close to the NOAEL found in some studies. However, it could be argued that BMD modelling may not be appropriate in the absence of an understanding of an endpoint.

436. In another approach the BMD related to a change in response equal to one standard deviation above the negative control mean has also been suggested (Crump 1984, 1995; Kavlock *et al*., 1995). The US EPA had suggested using a change equivalent to one standard deviation (1 SD) in the endpoint. A 5% decrease in BMR for foetal weight and a 10% decrease for brain cholinesterase have been suggested as adverse.

437. Another approach, based on change relative to the dynamic range (the maximum to minimum values which links to the Hill or exponential model parameters), has been suggested (Murrell *et al*., 1998). Change relative to the dynamic response may make comparison across endpoints possible. Other options include identifying doses where there are 'change points' on the dose response curve or the steepest point on the sigmoid curve.

438. One consideration is whether quantitative data should be converted into either binary or ordinal data so that it can be handled as if it were quantal data. Quantal data can in this approach be thought of as being on a continuous scale with various classes being defined as specific break points. Such dichotomization, however, results in a loss of information.

439. In conclusion, although modelling can appear a precise and a formal process, there is the opportunity for an appreciable amount of expert but subjective input into, for instance, the choice of options and the inclusion or exclusion of outliers and anomalous curve fits. Similar opportunities for expert but subjective judgement also apply to the NOAEL approach.

4.27 Other statistical methods for identifying change points in a dose-response relationship

440. Other statistical approaches have been developed for identifying "change points" in a dose-response study as an alternative to the NOAEL for continuous data. The change point is defined as the largest dose level which has the same response as in the negative control group. West & Kodell (2005) propose a method that investigates the profile of the least squares criterion over each of the intervals between the dose points in an experiment. They carried out simulation studies to show that the 95% lower confidence interval of the estimate of the change point had better statistical properties than the NOAEL. West & Kodell suggest linking the approach to the BMD methodology and that the method will need to be developed for a range of relevant change point model but note that conventional toxicology studies may have too few doses to estimate the parameters that explain more complex dose-response relationships.

441. The NOSTASOT method (no statistical significance of trend test) identifies the maximal dose which is not significantly different from the negative control group (Tukey *et al*., 1985). In general the NOSTASOT dose is higher than the no-effect

Appendix 1 Common statistical methods used in the analysis of data

The following glossary gives a brief description of the various tests in the flowchart (Figure 1) plus other tests that are commonly encountered in chronic studies. The definitions below are based upon a number of widely available glossaries and in particular Everitt (1995).

Tests for outliers

Dixon/Massey test: A test for outliers in a sample

Extreme Studentized Deviate (ESD) Statistic: A method used for identifying outliers; also known as Grubbs' test

Tests for non-normality

Chi-square test: A goodness of fit test that a set of data come from a hypothesized distribution such as the normal.

Kolmogorov–Smirnov one-sample test: A method that tests for goodness of fit of the data to a defined distribution

Shapiro–Wilk test: A method that tests that a set of random variable arise for a specified probability distribution used to test for departure from normality.

Tests for homogeneity of variance

Bartlett's test: A test for the equality (homogeneity) of the variances of a number of samples. The test is sensitive to departures from normality

Levene's test: A test for the equality (homogeneity) of the variances of a number of samples. The test is less sensitive to departures from normality

F test of variances: A test for a difference in the size of two variances.

Assumed normally distributed data

1. Overall tests

Analysis of variance (ANOVA): Statistical methodology which partition variability attributable to various causes. Family of modelling approaches simplest and commonest of which is the fixed effects one-way ANOVA which compares means across a set of samples.

Analysis of covariance (ANCOVA): Extension of ANOVA which allows for possible effects of covariates on endpoint in to effects of treatments which may reduce error mean squares associated with analysis.

Pearson's correlation coefficient: A test of the association between two variables

Linear regression: A test of the relationship between the two variables: one the independent like the dose the other the dependent i.e., the response. Used to examine trends in dose effects and to test the significance of the regression slopes.

2. Pair-wise comparisons

Duncan's multiple range test: A modified version of the Newman-Keuls multiple comparison test used to test for multiple comparisons when the initial ANOVA between groups is significant.

Dunnett's t-test : A multiple comparison test which compares each of a number of treatments to a single control.

Scheffe's test: A multiple comparison test with less power than Newman–Keuls multiple range test.

Williams' t-test: A multiple comparisons method for comparing each of a number of treatments with a single control.

Student's t-test: A number of different tests but here the independent two sample t-test assuming equal variance in the two groups and testing for a difference between two means.

Satterthwaite test: An alternative to the pooled-variance t test, and is used when the assumption that the two populations have equal variances seems unreasonable. It provides a t statistic that asymptotically (that is, as the sample sizes become large) approaches a t distribution, allowing for an approximate t test to be calculated when the population variances are not equal. Also known as Welch's test,

Fisher's least significant difference (LSD) test: A pair-wise test equivalent to the independent two-sample t-test except that the estimate of error is based upon the within group error of an ANOVA.

Tukey's Honest Significant (HSD) Difference test: A single step multiple comparison method used after the initial ANOVA between groups is significant.

Non-parametric procedures (percentage values, ranks, etc.)

Kendall's coefficient of rank correlation: A non-parametric test of the association between two variables based upon ranks

Pearson's rank correlation: Another non-parametric test of the association between two variables based upon ranks

Mann–Whitney U-test: A non-parametric alternative to the independent two-sample t-test. Also called the Wilcoxon Rank Sum Test.

Kolmogorov–Smirnov two-sample test: A distribution-free method that tests for any difference between two population distributions.

Wilcoxon signed-rank test: A non-parametric alternative to the paired t-test for matched or paired data.

Kruskal–Wallis ANOVA test: A distribution-free method that is the analogue of the one-way analysis of variance. Which tests whether the groups to be compared have the same population mean.

Jonckheere-Terpstra test: A test for detecting departures from independence where both the rows and columns of a contingency table have a natural order.

Distribution-free multiple comparisons tests

Dunn's test: A multiple comparison test based upon the Bonferroni test

Shirley's test: A non-parametric equivalent of Williams' test.

Quantal data (mortalities, pathology findings, etc.)

Fisher's exact test: Test for independence of two variables forming a 2 x 2 contingency table, based upon the hypergeometric distribution.

R x C chi-square test: A measure of association between the row and column classification or a r x c contingency table of variables

Litchfield & Wilcoxon test: Graphical method of probit analysis for calculation ED50 and confidence intervals

Cochran-Armitage linear trend test: Chi-square test for linear trend in counts and proportions.

Multivariate methods

Hotellings T^2: A generalization of the t-test to multivariate data

MANOVA: A multivariate analysis of variance to test the equality of the means of more than 2 populations.

Survival-adjusted procedures for analysis of carcinogenicity data

Log-rank test: Compares the survival distributions of two or more samples, sometimes called the Mantel-Cox test.

Peto analysis: A test in IARC monograph combining a life table test for fatal tumours with a prevalence analysis for incidental tumours.

Life table test: A survival adjusted test for fatal cancers or cancers with observable onset times.

Hoel–Walberg procedure: A survival adjusted test for incidental tumours. Also called the prevalence method.

Logistic regression: A form of regression analysis used when the response is binary: i.e., tumour/no tumour.

Poly-k test: A survival-adjusted Cochran-Armitage test for testing for a dose-related trend and/or a pair-wise difference in the incidence of tumours.

REFERENCES

Ahn, H., R.L. Kodell and H. Moon (2000), "Attribution of Tumor Lethality and Estimation of Time to Onset of Occult Tumors in the Absence of Cause-of-death Information", *Applied Statistics,* 49 157-169.

Bailer, A.J., R.B. Noble and M.W. Wheeler (2005), "Model Uncertainty and Risk Estimation for Quantal Responses", *Risk Analysis*, 25 291-299.

Bailer, A.J. and C.J. Portier (1988), "Effects of Treatment-induced Mortality and Tumor-induced Mortality on Tests for Carcinogenicity in Small Samples", *Biometrics,* 44 417-431.

Bannasch, P., R.A. Griesemer, F. Anders, B. Becker, J.R. Cabral, G.D. Porta, V.J. Feron, D. Henschler, N. Ito, R. Kroes, P.N. Magee, B. McKnight, U. Mohr, R. Montesano, N.P. Napalkov, S. Nesnow, A.E. Pegg, G.N. Rao, V.S. Turusov, J. Wahrdrendorf and J. Wilbourn (1986), *Long-term Assays for Carcinogenicity in Animals, in Long-Term and Short-Term Assays for Carcinogens: A Critical Appraisal*, Editors: Montesano, R., Bartsch, H., Vainio, H., Wilbourn, J. & Yamasaki, H. IARC Scientific Publications No. 83, Lyon, France.

Bieler, G.S. and R.L.Williams (1993), "Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity", *Biometrics,* 49 793-801.

Breslow, N. (1970), "A generalized Kruskal–Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship", *Biometrics*, 57 579–594.

Butterworth, B.E., O.B. Mathre, K.E. Ballinger and O. Adalsteinsson (2004), "Contamination is a Frequent Confounding Factor in Toxicology Studies with Anthraquinone and Related Compounds", *Int J Toxicol,* 23 335-344.

Cox, D.R. (1959), "The Analysis of Exponentially Distributed Life-times with Two Types of Failures", *Journal of Royal Statistical Society*, Series B, 21 4121-421.

Cox, D.R. (1972), "Regression Models and Life Tables (with Discussion)", *Journal of Royal Statistical Society*, Series B, 34, 187-220.

Crump, K.S. (1984), "A New Method for Determining Allowable Daily Intakes", *Fund Appl Toxicol*, 4 854-871.

Crump, K.S. (1995), "Calculation of Benchmark Doses from Continuous Data", *Risk Analysis,* 15 79-89.

Dekkers, S., C. de Heer and M. Rennen. (2001), "Critical Effect Sizes in Toxicological Risk Assessment: A Comprehensive and Critical Evaluation", *Environ Toxicol Pharmacol*, 10 33–52.

Dekkers, S., J. Telman, M.A.J. Rennen, M.J. Appel and C. de Heer (2006), "Within-animal Variation as an Indication of the Minimal Magnitude of the Critical Effect Size for Continuous Toxicological Parameters Applicable in the Benchmark Dose Approach", *Risk Analysis,* 26 867–880.

Dempster, A.P., M.R. Selwyn and B.J. Weeks (1983), "Combining Historical and Randomized Controls for Assessing Trends in Proportions", *Journal of the American Statistical Association*, 78 221-227.

Dinse, G.E. (1994), "A comparison of Tumor Incidence Analyses Applicable in Single-sacrifice Animal Experiments", *Statistics in Medicine,* 13 689-708.

Dinse, G.E. and J.K. Haseman (1986), "Logistic Regression Analysis of Incidental-tumor Data from Animal Carcinogenicity Experiments", *Fundamental and Applied Toxicology* 6 751-770.

Dunson, D. B. and G.E. Dinse (2001), "Bayesian Incidence Analysis of Animal Tumorigenicity Data", *Appl Stat,* 50 125–41.

Elmore, S. and S.D. Peddada (2009), "Points to Consider on the Statistical Analysis of Rodent Cancer Bioassay Data when Incorporating Historical Control Data", *Toxicologic Pathology,* 37 672-676.

EFSA (2005), *Opinion of the Scientific Committee on a request from EFSA related to a harmonised approach for risk assessment of substances which are both genotoxic and carcinogenic,* The EFSA Journal, 282 1–31.

EFSA (2009), *Guidance of the Scientific Committee on a request from EFSA on the use of the benchmark dose approach in risk assessment*, The EFSA Journal, 1150: 1-72.

EPA (1986), *Guidelines for Carcinogen Risk Assessment*, U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC.

EPA (2005), *Guidelines for Carcinogen Risk Assessment,* U.S. Environmental Protection Agency, Washington, DC.

Everitt, B.S. (1995), *The Cambridge Dictionary of Statistics in the Medical Sciences,* Cambridge University Press.

FDA Redbook (2007), *Toxicological Principles for the Safety Assessment of Food Ingredients*, Food and Drug Administration, Washington, DC.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, (3rd edition) Oliver and Boyd, Edinburgh.

Fisher, R.A. (1950), *Statistical Methods for Research Workers,* (11th edition) Oliver and Boyd, Edinburgh.

Fung, K.Y., D. Krewski and R.T. Smythe (1996), "A Comparison of Tests for Trend with Historical Controls in Carcinogen Bioassay", *Can. J. Statist.,* 24 431-454.

Gad, S.D. (2000), *Statistics for Toxicologists*, In Principles and Methods of Toxicology, Ed: A. Wallace Hayes, 4th edition, Taylor and Francis: Boston, MA, pp 369-451.

Gad, S. (2006), *Statistics and Experimental Design for Toxicologists and Pharmacologists*, 4th Edition, Taylor and Francis, Boca Raton, FL.

Gart, J.J., D. Krewski, P.N. Lee, R.E. Tarone and J. Wahrendorf (1986), *Statistical Methods in Cancer Research, III The Design and Analysis of Long-term Animal Experiments*, IARC Scientific Publications 79. vol. III pp. 1–219.

Gehan, E.A. (1965), *A Generalized Wilcoxon Test for Comparing K Samples Subject to Unequal Patterns of Censorship*, Biometrika, 52 203-223.

Graves, T.S. (1994), "A Comparison of Test Procedures for Multidose Animal Carcinogenicity Assays under Competing Risks", *Journal of Biopharmaceutical Statistics*, 4 289 - 320.

Haseman, J.K. (1983), "A Reexamination of False-positive Rates for Carcinogenesis Studies". *Fundamental and Applied Toxicology,* 3 334-339.

Haseman, J.K. (1984), "Statistical Issues in the Design, Analysis and Interpretation of Animal Carcinogenicity Studies", *Environmental Health Perspective*, 58 385-392.

Haseman, J.K., G. Hajian, K.S. Crump, M.R. Selwyn and K.E. Peace (1990), *Dual Control Groups in Rodent Carcinogenicity Studies,* In Statistical Issues in Drug Research and Development, Ed, Peace, K. E. Marcel Dekker, New York.

Haseman, J.K., J. Huff and G.A. Boorman (1984), "Use of Historical Control Data in Carcinogenicity Studies in Rodents", *Toxicologic Pathology,* 12 126-135.

Haseman, J.K., J.E. Huff, G.N. Rao and S.L. Eustis (1989), "Sources of Variability in Rodent Carcinogenicity Studies", *Fundam Appl Toxicol*, 12 793-804.

Haseman, J.K., J.S. Winbush, M.W. O'Donnell (1986), "Use of Dual Control Groups to Estimate False Positive Rates in Laboratory Animal Carcinogenicity Studies", *Fundamental and Applied Toxicology*, 7 573-584.

Hoel, D.G. and H.E. Walburg (1972), "Statistical Analysis of Survival Experiments", *J Natl Cancer Inst,* 49 361-372.

Iatropoulos, M.J. (1984), "Appropriateness of Methods for Slide Evaluation in the Practice of Toxicologic Pathology", *Toxicol Pathol*, 12 305-306.

Iatropoulos, M.J. (1988), *Society of Toxicologic Pathologists Position Paper: "Blinded" Microscopic Examination of Tissues from Toxicologic or Oncogenic Studies*, in Carcinogenicity, The Design, Analysis, and Interpretation of Tong-Term Animal Studies, edited by H.C. Grice and J.L. Ciminera, ILSI Monographs, Spring-Verlag, New York.

Ibrahim, J.G. and L.M. Ryan (1996), "Use of Historical Controls in Time-adjusted Trend Tests for Carcinogenicity", *Biometrics,* 52 1478-1485.

Ibrahim, J.G., L.M. Ryan and M.-H. Chen (1998), "Use of Historical Controls to Adjust for Covariates in Trend Tests for Binary Data", *J. Amer. Statist. Assoc*., 93 1282-1293.

IPCS (2009), *Principles for Modelling Dose–Response for the Risk Assessment of Chemicals,* Environmental Health Criteria 48, International Programme on Chemical Safety, World Health Organization, Geneva, Switzerland.

JECFA (2005), *Joint FAO/WHO Expert Committee on Food Additives Sixty-fourth Meeting, Rome - Summary and Conclusions*, FAO/WHO.

Kaplan, E.L. and P. Meier (1958), "Nonparametric Estimation from Incomplete Observations", *J. Am. Stat. Assoc.,* 53 457-481.

Kavlock, R. J., B. C. Allen, E. M. Faustman and C. A. Kimmel (1995), "Dose–response Assessments for Developmental Toxicity. IV. Benchmark Doses for Fetal Weight Changes", *Fundam. Appl. Toxicol.,* 26 211–222.

Keenan, K.P., P. Laroque, G.C. Ballam, K.A. Soper, R. Dixit, B.A. Mattson, S. Adams and J.B. Coleman (1996), "The Effects of Diet, ad libitum Overfeeding, and Moderate Dietary Restriction on the Rodent Bioassay: the Uncontrolled Variable in Safety Assessment", *Toxicologic Pathology,* 24 757-768.

Kodell, R.L, J.H. Farmer, D.W. Gaylor and A.M. Cameron (1982), "Influence of Cause-of-death Assignment on Time-to-tumor Analyses in Animal Carcinogenesis Studies", *J Natl Cancer Inst.,* 69 659-64.

Kodell, R.L., K.K. Lin, B.T Thorn and J.J Chen (2000), "Bioassays of Shortened Duration for Drugs: Statistical Implications", *Toxicological Sciences,* 55 415-432.

Lee, P.N., J.S. Fry, W.R. Fairweather, J.K. Haseman, R.L. Kodell, J.J. Chen, A.J. Roth, K. Soper and D. Morton (2002), "Current Issues: Statistical Methods for Carcinogenicity Studies" *Toxicological Pathology*, 30 403–414.

Lin, K.K. and M.W. Ali (1994), *Statistical Review and Evaluation of Animal Tumorigenicity Studies*, In Statistics in the Pharmaceutical Industry, 2nd Ed., Ed. Buncher, C.R & Tsay J.Z. Marcel Dekker, New York.

Lin, K.K. and M.A. Rahman (1998), "Overall False Positive Rates in Tests for Linear Trend in Tumor Incidence in Animal Carcinogenicity Studies of New Drugs (with discussions)", *Journal of Pharmaceutical Statistics*, 81 1-22.

Lin, K.K. (2010), *Carcinogenicity Studies of Pharmaceuticals*, in Encyclopedia of Biopharmaceutical Statistics, Third edition, Revised and Expanded, Volume 1, edited by Shein-Chung Chow, Informa Healthcare, New York and London.

McConnell, E.E., H.A. Solleveld, J.A. Swenberg and G.A. Boorman (1986), "Guidelines for Combining Neoplasms for Evaluation of Rodent Carcinogenesis Studies", *J. Natl. Cancer Inst.,* 76 283-289.

Murrell, J.A., C.J. Portier and R.W. Morris (1998), "Characterizing Dose-response I: Critical Assessment for the Benchmark Dose Concept", *Risk Anal*, 18 1326.

NTP (2005), *Description of Study Types: Statistical Procedures*, US National Toxicology Program (NTP) Website: http://www.niehs.ntp.gov.

OECD (2002) *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies,* Series on Testing and Assessment N° 35, OECD, Paris.

OECD (2009), *Carcinogenicity Studies*, OECD Guideline for the Testing of Chemicals, No. 451, OECD, Paris.

OECD (2009), *Chronic Toxicity Studies,* OECD Guideline for the Testing of Chemicals, No. 452, OECD, Paris.

OECD (2009), *Combined Chronic Toxicity/Carcinogenicity Studies,* OECD Guideline for the Testing of Chemicals, No. 453, OECD, Paris.

Peddada, S.D., G. Dinse and G. Kissling (2007), "Incorporating Historical Control Data when Comparing Tumor Incidence Rates", *J. Amer. Stat. Assoc.,* 102 1212-1220.

Peto R. (1974), "Guidelines on the Analysis of Tumour Rates and Death Rates in Experimental Animals", *British J Cancer*, 29 101-105.

Peto, R., M.C. Pike, N.E Day, R.G. Gray, P.N. Lee, S. Parish, J. Peto, S. Richards and J. Wahrendorf (1980), *Guidelines for Simple Sensitive Significance Tests for Carcinogenic Effects in Long-Term Animal Experiments,* In: IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,

supplement 2: Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal, Lyon: International Agency for Research on Cancer 311-346.

Portier, C.J., J.C. Hedges and D.G. Hoel (1986), "Age-specific Models of Mortality and Tumor Onset for Historical Control Animals in the National Toxicology Program's Carcinogenicity Experiments", *Cancer Res*, 46 4372−4378.

Portier, C. and D.G. Hoel (1983), "Optimal Bioassay Design Under the Armitage-Doll Multistage Model", *J Toxicol Environ Health,* 12 1−19.

Portier, C. and D.G. Hoel (1984), "Design of Animal Carcinogenicity Studies for Goodness-of-fit to Multistage Models", *Fundam Appl Toxicol.,* 4 949−959.

Prasse, K. (1986), Letter to the Editor (on blinded microscopic evaluation of slides from toxicity and carcinogenicity studies), *Toxicology and Applied Pharmacology,* 83 184-185.

Rahman, M.A. and K.K. Lin (2008), "A Comparison of False Positive Rates of Peto and Poly-3 Methods for Long-term Carcinogenicity Data analysis Using Multiple Comparison Adjustment Method Suggested by Lin and Rahman", *J Biopharm Statist*, 18 949-958.

Sand, S., K. Victorin and A.F. Filipsson (2008), "The Current State of Knowledge on the Use of the Benchmark Dose Concept in Risk, *J Appl Toxicol*., 28 405-21.

Slob, W. (2002), "Dose–response Modeling of Continuous Endpoints", *Toxicol. Sci.* 66 298–312.

Slob, W. and M.N. Pieters (1998), "A Probabilistic Approach for Deriving Acceptable Human Intake Limits and Human Health Risks from Toxicological Studies: General Framework", *Risk Analysis* 18 787–798.

Snedecor, G.W. and W.G. Cochran (1967), *Statistical Methods* (6th edition), The Iowa State University Press, Ames, Iowa.

Snedecor, G.W. and W.G. Cochran (1980), *Statistical Methods,* (7th edition), The Iowa State University Press, Ames, Iowa.

STP Working Group (2001), "Draft Recommendations on Classification of Rodent Neoplasms for Peto Analysis", *Toxicologic Pathology*, 29 265-268.

STP Peto Analysis Working Group (2002), "The Society of Toxicologic Pathology's Recommendations on Statistical Analysis of Rodent Carcinogenicity Studies", *Toxicologic Pathology,* 30 415–418.

Tarone, R.E. (1975), "Tests for Trend in Life Table Analysis", *Biometrika*, 62 679-682.

Tarone, R.E. (1982), "The Use of Historical Control Information in Testing for a Trend in Proportions" *Biometrics*, 38 215-220.

Tarone, R.E. and J. Ware (1977), "On distribution-free Tests for Equality of Survival Distributions", *Biometrika*, 64 165-60.

Temple, R.T., W.R. Fairweather, V.C. Glocklin, and R.T. O'Neill (1988), "The Case for Blinded Slide Reading", *Comments on Toxicology*, 2 99-109.

Thomas, D.G., N. Breslow and J.J. Gart (1977), "Trend and Homogeneity Analyses of Proportions and Life Table Data", *Computer and Biomedical Research*, 10 373-381.

Travis, K., I. Pate and Z. Welsh (2005), "The Role of the Benchmark Dose in a Regulatory Context", *Regulatory Toxicology and Pharmacology,* 43 280-291.

Tukey, J.W., J.L. Ciminera and J.F. Heyse (1985), "Testing the Statistical Certainty of a Response to Increasing Doses of a Drug", *Biometrics,* 41 295-301.

van Belle, G. (2009), *Statistical Rules of Thumb*, 2nd Edition, Wiley Hobeken.

WHO (1999), *Evaluation of Certain Food Additives and Contaminants,* Forty-Ninth Report of the Joint FAO/WHO Expert Committee on Food Additives, WHO Technical Report Series, No. 884, WHO, Geneva.

West, R.W. and R.L. Kodell (2005), "Changepoint Alternatives to the NOAEL", *Journal of Agricultural, Biological and Environmental Statistics*, 10 197-211.

Williams, D.A. (1971), "A test for Differences between Treatment Means when Several Dose Levels are Compared with a Zero Dose Control", *Biometrics*, 27 103-117.

Woutersen, R.A., D. Jonker, H. Stevenson, J.D. te Biesebeek and W. Slob (2001), "The Benchmark Dose Approach Applied to a 28-day Toxicity Study with Rhodorsil Silane in rats: the Impact of Increasing the Number of Dose Groups", *Food Chem. Toxicol*., 39 697-707.

5 DEFINITIONS/GLOSSARY

ADI/TDI: Acceptable daily intake/Tolerable daily intake: the amount of a test article in food or drinking water that can be ingested (orally) over a lifetime without an appreciable health risk.

Adverse Effect: Change in the morphology, physiology, growth, development, reproduction or life span of an organism, system, or (sub) population that results in an impairment of functional capacity, an impairment of the capacity to compensate for additional stress, or an increase in susceptibility to other influences.

AIC Akaike's Information Criteria

AUC: Area Under the Curve (Area under the plasma concentration-time curve): Area under the curve in a plot of concentration of substance in plasma over time.

Autolysis: The destruction of a cell through the action of its own enzymes.

Bayesian: Relating to statistical methods based on Bayes' theorem. A statistical approach that assesses the probability of a hypothesis being correct (for example, whether an association is valid) by incorporating the prior probability of the hypothesis and the experimental data supporting the hypothesis.

Benchmark dose (BMD): The dose corresponding to a small specified increase in effect over the background level.

BMDLx: is the lower confidence limit of BMD (see above) Typically, BMDL1 (with a response rate set at 1%) or BMDL10 (with a response rate set at 10%) is selected).

BMR: Benchmark Response

Bioaccumulation: the accumulation of the test article in an exposed organism. Bioaccumulation occurs when an organism absorbs a test article at a rate greater than that at which it is excreted.

Bias: A systematic error occurring in a measurement that is inherent in the sample itself or caused by the operator.

Bioavailability: Fraction of an administered dose that reaches the systemic circulation or is made available at the site of physiological activity.

Biomarker: A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Carcinogenicity: Substances are defined as carcinogenic if they induce tumours (benign or malignant), increase its incidence or shorten the time of tumour, when inhaled, ingested, dermally applied or injected.

CES: Critical Effect Size

Chronic Toxicity: Toxicity (adverse effect) after an exposure period of 12 months or longer due to a test article that has been ingested inhaled, dermally applied or injected.

COD: Cause of Death

COO: Context of Observation

Detoxification pathways: Series of steps leading to the elimination of toxic substances from the body, either by metabolic change or excretion.

Dose: Total amount of a test article administered to, taken up by, or absorbed by an organism, system, or (sub) population.

Dose-response: Relationship between the amount of an agent administered to, taken up by, or absorbed by an organism, system, or (sub) population and the change developed in that organism, system, or (sub) population in reaction to the agent.

Dyschromatopsia: Deficiencies of colour vision

Dysplasia: refers to any disordered growth and maturation of an epithelium.

Electron microscope: A type of microscope that uses a beam of electrons, rather than light to produce an image of a specimen for detailed observation.

Extrapolation: Inference of one or more unknown values on the basis of that which is known or has been observed.

Exposure: Concentration or amount of the test article that reaches a target organism, system, or (sub) population in a specific frequency for a defined duration.

Fixation: A chemical process by which biological tissues are preserved from decay, to as close to its natural state as possible in the process of preparing tissue for examination.

Frequentist: An advocate of frequency probability. This is the inference framework in which the well-established methodologies of statistical hypothesis testing and confidence intervals are based

Fundoscopy (See ophthalmascopy)

Genomic(s): The study of all of the nucleotide sequences, including structural genes, regulatory sequences, and noncoding DNA segments, in the chromosomes of an organism.

Genotoxic/genotoxicity: A deleterious action on a cell's genetic material affecting its integrity.

GLP: Good Laboratory Practice

Hazard: The inherent property of a test article to cause adverse effects when an organism, system, or (sub) population is exposed to that test article.

Hazard identification: The identification of the type and nature of adverse effects that an agent has as inherent capacity to cause in an organism, system or (sub) population.

Haematology: The study of blood, the blood-forming organs, and blood diseases.

Histochemistry: The branch of histology dealing with the identification of chemical components in cells and tissues.

Histology: The study of the microscopic anatomy of cells and tissues of animals.

Histopathology: The study of the microscopic anatomical changes in diseased tissue

Hyperplasia: The proliferation of cells within an organ or tissue beyond that which is ordinarily seen.

Induction/Enzyme induction: Enzyme synthesis in response to an environmental stimulus or inducer molecule;

Immunohistochemistry: The process of localizing antigens (e.g., proteins) in cells of a tissue section exploiting the principle of antibodies binding specifically to antigens in biological tissues.

Local effect: Adverse effect at the site of first contact (e.g., skin, eye, mucous membrane/gastro-intestinal tract, or mucous membrane/respiratory tract).

LO(A)EL, also **LOAEL**: Lowest Observed (Adverse) Effect Level: The lowest level of a test substance that causes an observed and significant adverse effect on the test species compared with the controls.

LO(A)EC, **also LOAEC**: Lowest Observed (Adverse) Effect Concentration: The lowest concentration of a test article that causea an observed and significant adverse effect on the test species compared with the controls.

LOEC: Lowest Observed Effect Concentration. The lowest concentration of a test article that causes an observed and significant effect on the test species compared with the controls.

LOEL: Lowest Observed Effect Level: The lowest level of a test substance that causes an observed and significant effect on the test species compared with the controls.

Macroscopic/macroscopy: An observation that is large enough to be perceived or examined by the unaided eye.

Mechanism of Action (MOA): The individual biochemical and physiological events leading to a toxic effect).

Metabolism: In the context of this Guidance, refers to all the chemical processes in the body that alter the structure of the substanve administered.

Microscopic/microscopy: An observation of something extremely small in size; visible only with the aid of a microscope.

Mode of Action (MOA): Mode of Action: the processes by which a chemical induces toxicity. A MOA can inform about relevance of observed effects in laboratory animals to humans and the variability of response within the human population.

Morphology: The form and structure of organs and tissues and their specific structural features.

MTD: Maximum Tolerated Dose

MTC: Maximum Tolerated Concentration

Mydryatic agent: A chemical agent that induces dilation of the pupil.

Neoplasia: Refers to tumour growth. A neoplasm may be benign or malignant.

NO(A)EC, also **NOAEC**: No-Observed- Adverse-Effect-Concentration. The highest concentration of a test substance that does not cause any observed and statistically significant adverse effect on the test species compared with the controls.

NO(A)EL, also **NOAEL**: No-Observed- Adverse-Effect-Level. The highest level of a test substance that does not cause any observed and statistically significant adverse effect on the test species compared with the controls.

NOEC: No-Observed-Effect-Concentration. The highest concentration of a test substance that does not cause any observed and statistically significant effect on the test species compared with the controls.

NOEL: No-Observed-Effect-Level. The highest level of a test substance that does not cause any observed and statistically significant effect on the test species compared with the controls.

NOSTASOT: No statistical significance of trend (test)

Ocular: Relating to the eye or the sense of sight.

Ocular adnexa: Structures adjacent to the eye such as the lacrimal apparatus, the extraocular muscles and the eyelids, and the conjunctiva

Ophthalmoscopy (funduscopy or fundoscopy): Visual non-invasive examination the interior of the eye, including the lens, retina and optic nerve using a specialised instrument (ophthalmoscope or funduscope) containing a concave mirror and a battery-powered light

Palatability/ decrease in palatability: Acceptable to the taste, sufficient agreeable to be eaten/reduced taste / makes the food less aggreable due to organoleptic changes e.g., change in odor/flavor leading to reduced food intake.

Pathology: The study of disease.

Perimortem: At or around the time of death.

POD: Point of Departure: The dose-response point that marks the beginning of a low-dose extrapolation. This point is most often the upper bound on an observed incidence or on an estimated incidence from a dose-response model.

Postmortem: Done, occurring, or collected after death

(Q)SAR: Quantitative Structure Activity Relationship

Randomisation: A term used to describe the selection of samples for each 'arm' of a study or experiment based on chance alone– i.e., a theoretical coin toss, which is intended to minimize the influence of irrelevant details and selection bias, and produce statistically valid data.

Read-across: The endpoint information for one or more chemicals is used to make prediction of the endpoint for the target chemical i.e., use of structure similarity analysis to identify analogous compounds to determine whether there may be opportunities to bridge databases for one chemically structurally similar compound to another.

Reference dose (RfD): An estimate of a daily exposure to a chemical that is unlikely to cause harmful effects during a lifetime.

RP: Reference Point

Route of administration (oral, IV, dermal, inhalation, etc.): Refers to the means by which substances are administered to the body (e.g., orally by gavage, orally by diet, dermal, by inhalation, intravenously, etc).

Route-to-route extrapolation: The prediction of an equivalent dose and dosing regime that produces the same toxic endpoint or response as that obtained for a given dose and dosing regime by another route.

SF: Safety Factor

SAR: Structure Activity Relationship

Systemic effect: A toxicological effect that affects the entire body or many organs.

Systems Modeling: (Pharmacokinetic-based, Physiologically-based Pharmacokinetic, Biologically-based, etc.): Abstract model that uses mathematical language to describe the behaviour of a system.

Target tissue: Tissue in which the principal adverse effect of a toxicant is manifested.

TDI: Tolerable Daily Intake

Threshold: Dose or exposure concentration of an agent below which a stated effect is not observed or expected to occur.

Toxicity: Inherent property of an agent to cause an adverse biological effect.

Toxicodynamics: The processes of interaction of toxicologically active substances with target sites, and the biochemical and physiological consequences leading to adverse effects i.e., the way in which the chemical substance behaves/interacts within the biological system to cause toxicity

Toxicokinetics (Pharmacokinetics): A term describing the processes of chemical absorption, distribution, metabolism, and excretion in the organism (ADME) i.e., the way in which the chemical substance is absorbed, moves within the body and are being excreted.

Tonometry: The measurement of tension or pressure. In ophthalmoscopy, tonometry measures intraocular pressure by recording the resistance of the cornea to pressure (indentation).

Transcriptomics: The study of the complete set of RNA transcripts produced by the genome at any one time.

UF: Uncertainty Factor

VSD: Virtually Safe Dose