From INSTITUTE OF ENVIRONMENTAL MEDICINE
Karolinska Institutet, Stockholm, Sweden

# WORTH WEIGHTING FOR - STUDIES ON BENCHMARK DOSE ANALYSIS IN RELATION TO ANIMAL ETHICS IN TOXICITY TESTING

Joakim Ringblom

# WORTH WEIGHTING FOR - STUDIES ON BENCHMARK DOSE ANALYSIS IN RELATION TO ANIMAL ETHICS IN TOXICITY TESTING THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

## Joakim Ringblom

*Principal Supervisor:*
Associate Professor Mattias Öberg
Karolinska Institutet
Institute of Environmental Medicine
Unit of Work Environment Toxicology
and Swedish Toxicology Sciences Research
Center, Swetox

*Co-supervisor:*
Professor Gunnar Johanson
Karolinska Institutet
Institute of Environmental Medicine
Unit of Work Environment Toxicology

*Opponent:*
PhD Harvey Clewell III
ScitoVation

*Examination Board:*
Professor Anders Bignert
Swedish Museum of Natural History
Department of Environmental Research and
Monitoring

Professor Magnus Breitholtz
Stockholm University
Department of Environmental Science and
Analytical Chemistry

Associate Professor Helena Röcklinsberg
Swedish University of Agriculture
Department of Animal Environment and Health
Division of Environment, Care and Herd Health

# ABSTRACT

A purpose of chemical health risk assessment is to characterize the nature and size of the health risk associated with exposure to chemicals, including identification of a dose below which toxic effects are not expected or negligible. This is usually based on analysis of dose-response data from toxicity studies on animals. Traditionally the dose-response in animals has been analyzed employing the No-Observed-Adverse-Effect-Level (NOAEL) approach, but because of the several flaws of this approach it is to a greater and greater extent being replaced by the so called Benchmark Dose (BMD) approach.

Previous evaluations of how to design studies in order to obtain as much information as possible from a limited number of experimental animals have revealed the importance of including high doses. However, these studies have not taken the distress of the laboratory animals, which is likely to be higher at high doses, into account.

The overall aim of the present thesis was to examine how study designs, especially with dose groups of unequal size, affect the quality of BMD estimates and level of animal distress.

In **Paper I** our computer simulations concerning the appropriateness of using nested models in BMD modelling of continuous endpoints indicate that it is problematic to calculate BMD on the basis of simpler models and that they should be used with caution in connection with risk assessment as they may result in underestimations of the true BMD.

In **Paper II-III** our computer simulations of toxicity testing with unequal group sizes showed that better information about dose-response can be obtained with designs that also reduce the level of animal distress.

In **Paper IV** we interviewed members of the Swedish Animal Ethics Committees concerning how the number of animals used in toxicity tests might be weight against the distress of the individual animal. Their opinions concerning whether it is preferable to use fewer animals that suffer more rather than a large number of animals that suffer a little, differed considerably between individuals. However, there were no statistically significant differences in relation to the fact that respondent were either researchers, political representatives or representatives of animal welfare organizations.

In **Paper V** the results from **Paper IV** and the simulation techniques in **Paper II** were combined to evaluate how toxicity tests could be designed to obtain as much information as possible at a limited ethical cost, with respect to both the number of animals used and their individual distress. The most ethically efficient design depended on what constituted the ethical cost and how large that ethical cost was.

In conclusion, this thesis describes the potential to use BMD-aligned study design as a mean for refinement of animal toxicity testing. In addition, new strategies for model selection and quantitative measures of ethical weights are presented.

# POPULÄRVETENSKAPLIG SAMMANFATTNING

Vi utsätts ständigt i olika grad för kemikalier som potentiellt kan vara skadliga för oss. Hälsoriskbedömningar av dessa kemikalier görs för att avgöra när åtgärder behöver sättas in för att begränsa vår exponering. Riskbedömningarna mynnar ofta ut i sättandet av riktvärden, såsom acceptabla dagliga intag. Riktvärdena baseras ofta på djurförsök där man identifierar NOAEL-värden, dvs. den högsta dos som inte ger en statistiskt säkerställd effekt. NOAEL metodiken har dock brister och därför använder allt fler Benchmark Dos (BMD) metoden. En fördel med BMD-metoden är att den tar hänsyn till osäkerheter i data på ett bättre sätt.

Flera tidigare undersökningar har studerat hur man designar ett toxikologiskt försök för att få ut så mycket information som möjligt. Dessa studier har utgått ifrån ett bestämt totalt antal försöksdjur och bland annat visat att det är viktigt att det förekommer höga doser i försöken. Dock har inga tagit hänsyn till de etiska aspekterna av försöket. I samband med toxikologiska tester är det till exempel rimligt att tänka sig att djur som utsätts för en hög dos lider mer än djur som får en lägre dos.

Det övergripande målet var därför att studera hur designen av toxikologiska försök kan förbättras så att likvärdig, eller bättre, information kan tas fram med lika mycket eller mindre lidande hos försöksdjuren.

Under arbetet med våra datasimuleringar observerade vi att den vedertagna BMD-metodiken i vissa fall kan leda till värden som underskattar risken. I **studie I** undersökte vi varför, hur ofta och när detta inträffar. Vi visade att fenomenet uppkommer när alltför enkla matematiska modeller väljs för att beskriva sambandet mellan dos och effekt.

I **studie II-III** har vi med datorsimuleringar undersökt förhållandet mellan kvalitén på den information man får från ett toxikologiskt och hur försöket läggs upp. Vi visade att i flera fall kan kvaliteten på denna information förbättras samtidigt som djurlidandet minskas.

I **studie IV** intervjuade vi ledamöterna i de svenska djurförsöksetiska nämnderna om hur tecken på djurs lidande bör värderas etiskt. Vilket är det minst dåliga alternativet; att ha ett fåtal djur som utsätts för ett större lidande eller att använda fler djur som lider mindre? Individuella ledamöter resonerade väldigt olika kring dessa frågor. Vi såg dock inga säkra skillnader mellan forskare, politiker eller representanter från djurskyddsorganisationer.

I **studie V** kombinerade vi resultaten från studie IV med datorsimuleringarna från **studie II** för att undersöka hur ett toxikologiskt försök skulle kunna läggas upp för att få ut mest information givet en fast etisk kostnad för försöket, där den etiska kostnaden var beroende både av antalet djur i försöket och av djurens lidande. Det visade sig att det finns potentiella etiska vinster i förändrad studiedesign, beroende på hur den etiska kostnaden definieras.

Sammantaget visar avhandlingen att det finns en potential att minska djurs lidande utan att förlora vetenskaplig information genom att använda moderna metoder för dos-responsanalys samt att djurförsök kan värderas både utifrån antalet djur och utifrån deras lidande.

# LIST OF SCIENTIFIC PAPERS

I. **Ringblom J**, Johanson G, Öberg M. Current Modeling Practice May Lead to Falsely High Benchmark Dose Estimates. Regulatory Toxicology and Pharmacology. 2014 Jul;69(2):171-7

II. Kalantari F\*, **Ringblom J**\*, Sand S, Öberg M. Influence of Distribution of Animals between Dose Groups on Estimated Benchmark Dose and Animal Distress for Quantal Responses. Accepted for publication by Risk Analysis.

III. **Ringblom J**, Kalantari F, Johanson G, Öberg M. Influence of Distribution of Animals between Dose Groups on Estimated Benchmark Dose and Animal Distress for Continuous Effects. Manuscript.

IV. **Ringblom J**, Törnqvist E, Hansson S-O, Rudén C, Öberg M. Assigning ethical weights to clinical signs observed during toxicity testing. ALTEX. 2016 Jul 21. doi: 10.14573/altex.1512211. [Epub ahead of print]

V. **Ringblom J**, Johanson G, Öberg M. Dose-Response Simulations Suggest that Toxicity Tests may be Optimized in Relation to the Ethical Cost. Manuscript.

\*=These authors contributed equally to this work.

# CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 3R | Replacement, Reduction, Refinement. |
| AEC | Animal Ethics Committee |
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |
| BMA | Bayesian Model Averaging |
| BMD | Benchmark Dose |
| BMDL | Lower limit of a confidence interval for the BMD |
| BMDU | Upper limit of a confidence interval for the BMD |
| BMR | Benchmark Response |
| CES | Critical Effect Size |
| CV | Coefficient of Variation |
| EFSA | European Food Safety Authority |
| KIC | Kullback Information Criterion |
| NOAEL | No-Observed-Adverse-Effect-Level |
| NTP | National Toxicology Program |
| PoD | Point of Departure |
| PTO | Person Trade-Off |
| QALY | Quality Adjusted Life Years |
| RC | Repugnant Conclusion |
| REACH | EU legislation concerning Registration, Evaluation, Authorisation and Restriction of Chemicals |
| RfC | Reference Concentration |
| RfD | Reference Dose |
| RMSE | Root Mean Squared Error |
| UF | Uncertainty Factor |
| USAWA | United States Animal Welfare Act |
| USEPA | United States Environment Protection Agency |

# 1   INTRODUCTION

Current toxicity testing involves use of many non-human animals (hereafter simply referred to as just "animals"), which has been criticized on both ethical (Regan, 1983; Singer, 2009) and scientific grounds (see Knight, 2013 for a review). This thesis does not resolve such controversies, but focuses on how to use animals as efficiently as possible, from a mathematical and statistical perspective. In addition, the ethical aspects of animal experiments are introduced as a limiting factor in optimization of experimental design. Many of the conclusions in **Papers I-III** are also relevant for *in vitro* toxicity testing.

## 1.1 RISK ASSESSMENT

An aim of quantitative chemical risk assessment is to assess the risk associated with the chemical exposure in a target population such as workers or the general public. The risk assessment process consists of 4 different steps (Figure 1) (NRC, 1983; WHO/IPCS, 2004):
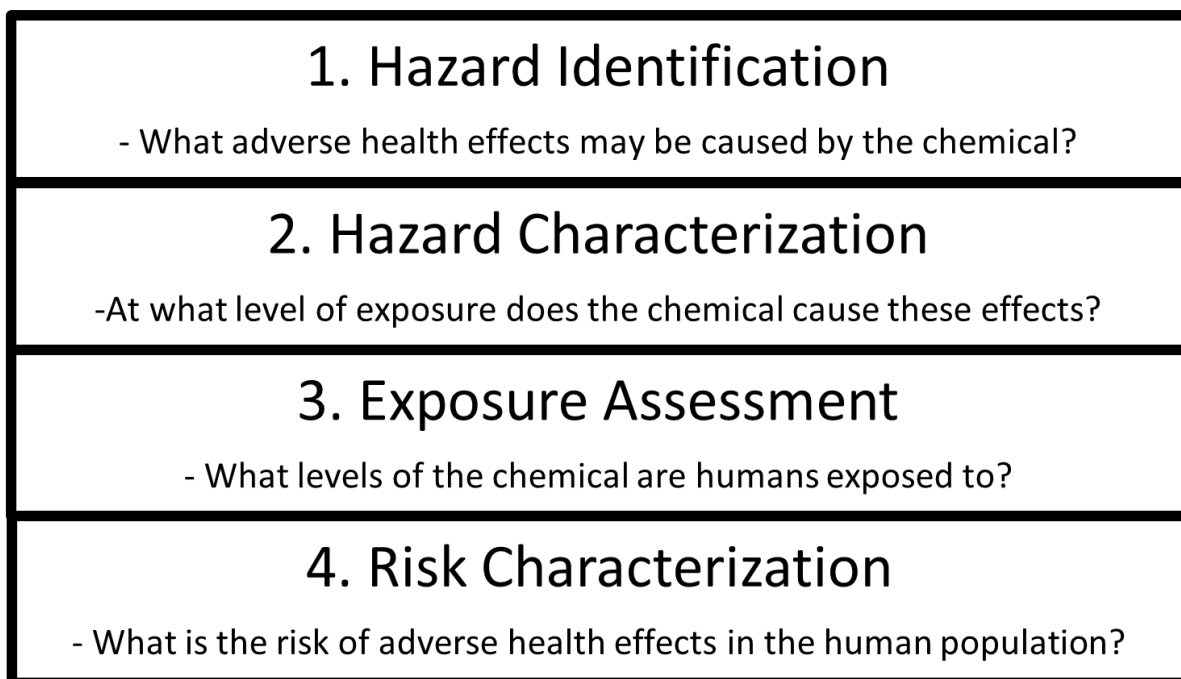
## 1. Hazard Identification
- What adverse health effects may be caused by the chemical?

## 2. Hazard Characterization
-At what level of exposure does the chemical cause these effects?

## 3. Exposure Assessment
- What levels of the chemical are humans exposed to?

## 4. Risk Characterization
- What is the risk of adverse health effects in the human population?

**Figure 1. The four parts of chemical risk assessment.**

During the exposure characterization, the exposure to chemical through various sources is estimated. Chemicals can be taken up via the gastrointestinal tract, via inhalation or through the skin and the sources of exposure varies between different chemicals and human populations. This includes exposures via food, drinks, exposures in the workplace or as exposures as results of accidents etc. Finally, at the risk characterization stage all evidence from the previous steps is weight together to determine the human risk associated with the exposure.

The RfDs being determined in the dose-response assessment is derived from so called Points-of-Departure (PoDs), i.e doses that exert no or acceptably low adverse effect in the study of interest. To obtain the RfD, the PoD is divided by a number of different Assessment Factors (AF) that take into account the uncertainties resulting from extrapolation of animal data to humans, as well as other uncertainties:

$$\text{RfD} = \frac{PoD}{AF} \quad \text{(US EPA, 2002)}$$

The traditional AF of 100 consists of two parts, a factor of 10 designed to take into account the toxicokinetic and toxicodynamic differences between the species tested and humans, along with a factor 10 that reflects differences in sensitivity between different human individual (ECHA, 2012). Additional assessment factors may also be applied when extrapolating from short- to a long-term exposure and/or when utilizing incomplete databases. When quantitative information concerning the difference in sensitivity between animals and humans or between different individuals are available, chemical specific assessment factors should be used instead of the standard assessment factors (Meek et al., 2002; US EPA, 2002).

For most chemicals, it is assumed that there is a threshold level of exposure below which there is no risk of adverse effects (Dybing et al., 2002; Edler et al., 2002), but it is extremely difficult or even impossible to determine the existence of such a threshold based on experimental data (Slob, 1999; Slob, 2007). For genotoxic carcinogens it is however generally assumed that there is no threshold for the risk because a single genotoxic molecule could interfere with DNA leading to a mutation and cancer (US. EPA, 2005). This difference in the risk assessment of genotoxic carcinogens and other toxicants has its origins in the late 1970´s, when similarities between the effects of genotoxic carcinogens and radiation were realized (Bogdanffy et al., 2001).

### 1.1.1  No-Observed-Adverse-Effect-Level

Traditionally, the so-called No-Observed-Adverse-Effect-Level (NOAEL) has been used as the PoD. The NOAEL is the highest dose in a study that does not give rise to an adverse effect that is statistically significantly different from the effect in the control group (WHO, 1999). Since it is occasionally very difficult to determine whether an effect is actually adverse and relevant to humans the term No-Observed-Effect-Level (NOEL) is sometimes preferred (Berry, 1988). The term NOAEL will however be used throughout this thesis.  In studies where there is no NOAEL as the effect in all dose groups differs from the control, the RfD can be calculated from the Lowest Adverse Effect-Level (LOAEL) instead of the NOAEL, usually with application of an additional assessment factor of 3-10 (ECHA, 2012) .

### 1.1.2  Benchmark Dose

The Benchmark dose (BMD) approach (Figure 2) was introduced by Crump to circumvent some of the disadvantages connected with the use of the NOAELs to set RfDs (Crump, 1984). BMD was originally mostly used with quantal data from developmental studies (Allen

et al., 1994a; Allen et al., 1994b), but is now used for both other types of experimental data and epidemiological data (Budtz-Jorgensen et al., 2001; Sand et al., 2008).

Determination of a BMD involves first fitting a dose-response model to the data and then interpolating to find which dose that causes a predefined response. That dose is defined as the BMD. To account for uncertainty and provide a margin of safety, a two-sided 90% confidence interval for the BMD is calculated and the lower limit of that interval, the BMDL, is employed instead of the NOAEL to calculate RfDs. The upper limit of this confidence interval, the BMDU, is sometimes used to calculate the BMDU/BMDL ratio which provides an estimate of the uncertainty in the BMD value. The BMD/BMDL ratio can also be used for this purpose but is less good as it is does not take the full uncertainty in the BMD estimation into account (Slob, 2014a). The profile likelihood procedure (Venzon and Moolgavkar, 1988), which is relatively rapid, is commonly used to calculate the BMDL and BMDU, but other approaches have been discussed and used as well, such as the bootstrap and the delta methods (Moerbeek et al., 2004).

The BMD approach has numerous advantages over the usage of the NOAEL (Crump, 1984; Davis et al., 2011). The most important, of which is that BMD takes uncertainty into account in a proper manner. In a test with smaller dose groups, the NOAEL tends to be higher, giving rise to higher RfDs when there is not much data available. This is not reasonable from a precautionary perspective. With BMD on the other hand, the use of small groups usually results in lower, more precautionary, RfDs.
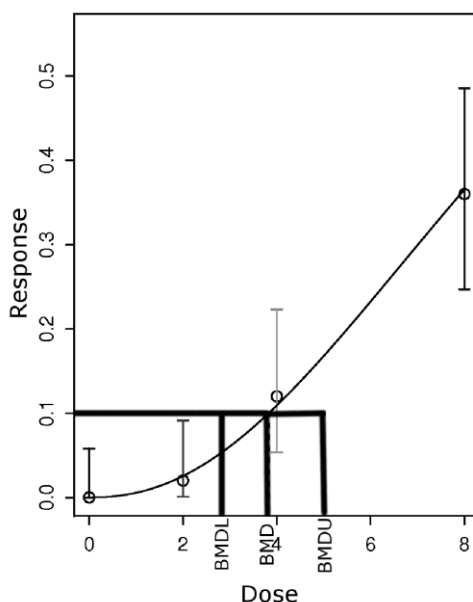


**Figure 2. The BMD procedure. After fitting a mathematical model to the dose-response data the BMD is the dose that causes a predefined response. The BMDL and BMDU values are the lower and upper limits of the 90% confidence interval for the BMD.**

Another advantage of BMD modeling is that it well suited to handle covariates, such as sex (Edler, 2014). Data from females and males can be used simultaneously in the curve fitting and the parameters of interest will only be covariate dependent if there is a difference between the sexes for those parameters. For the parameters where there is no difference, the different groups will share the same parameter, thereby extracting more information. Additional differences between the BMD approach and the NOAEL approach are listed in Table 1. Despite some reluctance to use the BMD approach (Travis et al., 2005), this method has now been implemented as an alternative or preferred approach by many regulatory agencies (Brandon et al., 2013; ECHA, 2012; NAC/AEGL, 2001; Solecki et al., 2005; USEPA, 1995; WHO, 2009).

**Table 1. Comparisons of the BMD and NOAEL approaches (Öberg, 2010; Slob, 2014a; Travis et al., 2005)**

| Advantages of the BMD | Disadvantages of the BMD |
|---|---|
| Takes uncertainty into account in a proper manner. | More difficult to perform. |
| More suitable for simultaneous analysis and pooling of datasets. | Less intuitive. |
| Promotes good quality experiments. | Less well known. |
| Takes the shape of the dose-effect curve into account. | Requires greater harmonization and consensus regarding the choice of models, benchmark responses etc. |
| The choice of critical effect size can reflect the severity of the effect. | |
| Less dependent on study design. | |
| Partially solves the "LOAEL only" problem. | |
| Set on a continuous scale. | |
| BMD ratios are more informative than NOAEL ratios. | |

### 1.1.2.1 Quantal data

With quantal data, also referred to as dichotomous or dose-response data, the outcomes are incidences, e.g. number of animals with tumors. Since each animal/human/cell either responds or not, quantal data lie between 0% and 100%. With such data the BMD is defined as the dose that gives rise to a Benchmark Response (BMR), most often defined as either an increased additional risk or extra risk:

$$Additional\ Risk = P(x) - P(0)$$

$$Extra\ Risk = \frac{P(x) - P(0)}{1 - P(0)}$$

An extra risk of 10% is recommended as default for the BMR by both EFSA (EFSA, 2009) and US EPA (US EPA, 2012). One advantage of using extra rather than additional risk is that a BMD based on extra risk and calculated with a multivariate method, will always be lower than the corresponding value calculated from each endpoint separately (Gaylor et al., 1998). It has also been suggested that the BMR could be defined as the effect at the Signal-to-Noise-Cross-over-Dose, i.e. a dose where the extra risk is equal to the background noise (Sand et al., 2011).

A multitude of dose-response models have been used for quantal data (Sand et al., 2008). Since the results of developmental toxicity studies are a special kind of quantal data the pups from the same litter are correlated. BMD modeling of developmental toxicity data therefore uses special types of models to take intra-literal effects into account (Kodell et al., 1991; Rai and Vanryzin, 1985).

### 1.1.2.2 Continuous data

Body weight, organ weights and enzyme levels are typical continuous data, also referred to as dose-effect data. For such data each animal has its own magnitude of effect and the arithmetic or geometric means of the different dose groups are usually compared. One important difference between quantal and continuous data is the inherent presence of an upper limit of 100% in the case of the former. Although most continuous effects have a lower and upper limit, the value of it is not known beforehand.

Originally continuous data were often modelled using linear models, power models or polynomials (Allen et al., 1994a; Allen et al., 1996; Crump, 1984; Kavlock et al., 1995). However, such models do not level off at higher doses and are thereby clearly not suitable for some datasets. As a consequence, there has been a stronger focus on models that do have the ability to level off at higher doses (US EPA, 2012), such as the Hill model (Barton et al., 1998; Murrell et al., 1998) and exponential model (Slob, 2002). Both the Hill model and the exponential model can be parametrized as families of nested models, i.e. the simpler models within a family can be derived from the more complex ones by fixing parameters in the latter (Figure 3).
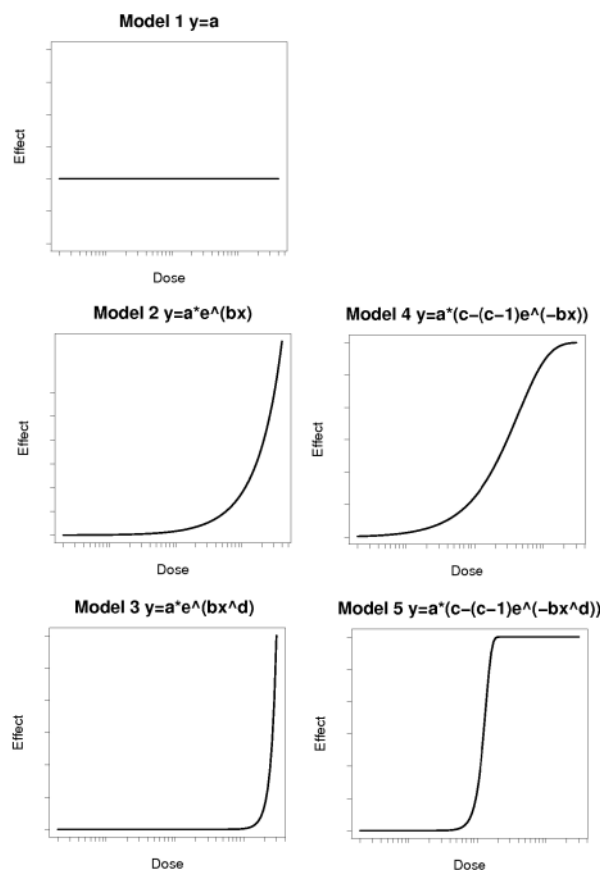
**Figure 3. The nested set of exponential models (Slob, 2002). Model 1 is obtained from model 2 by setting b = 0 and model 2 from model 3, 4 or 5 by setting c = 0 and/or d = 1. A more complicated model is selected if the fit is significantly better according to a likelihood ratio test. This image is taken from Ringblom et al (2014)**

When Slob and Setzer (2014) analyzed a large set of historical data sets including both *in vivo* and *in vitro* endpoints, they found that both the 4-parameter exponential and Hill models fitted the data adequately. Fitting the data simultaneously for the same endpoint, but different chemicals, gave curves of the same shape for the *in vitro* endpoints, as well as dose-effect curves of similar shape for the *in vivo* endpoints.

The different approaches to choosing a BMR for continuous data can be categorized into two categories, nonprobabilistic and probabilistic. The original definition was a nonprobabilistic definition with the BMR, or cBMR, defined as a percentage change in the mean effect compared to the mean background effect:

$$cBMR = \frac{m(x) - m(0)}{m(0)}$$

The cBMR was later renamed Critical Effect Size (CES) and discussed further by Slob and Pieters (Slob and Pieters, 1998). It is has been recommended that the CES should be a low but still measurable effect. Having a CES that is too low will lead to extrapolations and heavy dependence of the BMD on the model employed (Edler, 2014). Different endpoints therefore

requires different CES values (Dekkers et al., 2001). EFSA has proposed a preferred default 5% as a CES, with modifications if required by toxicological or statistical considerations (EFSA, 2009). It has also been suggested that CES values can be set by observing the intra-animal variation in historical data (Dekkers et al., 2006).

Other definitions of the BMR for continuous data have been proposed, e.g. as the effect corresponding to a percentage of the entire dose-effect span, so that a $BMD_{10}$ would be equivalent to a $ED_{10}$ (Murrell et al., 1998) or as the dose at that corresponds to where the slope of the dose-effect curve changes most rapidly (Sand et al., 2006).

The simplest probabilistic approach to define a BMR for continuous data is to transform the continuous data into quantal data, e.g. by defining a cutoff point such as weight loss of 5% or 10% as adverse. Any animal with a larger weight loss will be considered a responder and animals with less weight loss will be considered a non-responder. This procedure has however been criticized since information is lost when the data are quantalized (Crump, 1995; West and Kodell, 1999).

The probabilistic hybrid approach proposed by Gaylor and Slikker (1990) is more advanced. Here, the distribution of the effect at each dose level is estimated and the BMD defined as the dose that causes a predefined fraction of the animals to exhibit effects greater than a certain cut-off level. The U.S EPA supports the use of this procedure as the default approach to selecting a BMR, but only if there is no specific change in endpoint that can be considered adverse (US EPA, 2012).

Another important difference between quantal and continuous data is that for continuous data is that the latter requires assumptions concerning the distribution of the data (normally distributed, lognormally distributed or some other form of distribution). Shao and Small (2013) concluded that incorrect assumptions regarding normality or lognormality exert only minor impact on the BMD estimate when the variation is small. The variance of continuous data may also be assumed to be consistent at all doses or to change with the effect size. None of these assumptions are necessary with quantal data.

### 1.1.3 Choice of models and model-averaging techniques.

Application of different models to the same data will yield different values for the BMD and BMDL. As a consequence, there are different methods that guide the choice of which BMD and BMDL to use. The different methods rely on the goodness of fit of the model, often assessed as the loglikelihood of the fit. An acceptable fit can be examined by comparisons to the fit of the full saturated model and/or the fit of the reduced straight line model (Edler, 2014). While larger and more complex models usually provide better fits they may still be less preferable than simpler based on the principle of parsimony.

In the case of two nested models, statistical theory states that the difference in loglikelihood follows a chi-square distribution. This means that in a nested set of models, as with the family of exponential models or Hill models, the choice of model can be based on a series of

likelihood ratio tests (Figure 3). If addition of a parameter to a simple model does not significantly improve the fit the simpler model is retained. Following this procedure for all model comparisons in a nested set, a single model can be selected.

Current EFSA guidelines suggest that the lowest BMDL among the models that pass a goodness-of-fit test should be used as the PoD (EFSA, 2009). EPA´s guidelines are less conservative, suggesting that the model with the lowest AIC (Akaike Information Criterion) should be used as the PoD, unless there is a large difference between the BMDL values obtained with the different models (US EPA, 2012). The AIC takes the likelihood (L) of the model fit into account, but penalizes models with many parameters (k):

$$AIC = 2k - 2ln(L)$$

Accordingly, the EPA guidelines result in less conservative estimates of BMDLs than do the EFSA guidelines, which on the other hand can be seen as overly conservative.

Model averaging is a more advanced alternative which takes model uncertainty into consideration by weighting the contribution of various models together. The information from all of the models rather than only the most conservative or the one with the best fit is used to determine the PoD. Bayesian Model Averaging (BMA) has been used in to calculate BMDs in several BMD investigations (Dankovic et al., 2007; Morales et al., 2006; Shao and Gift, 2014; Simmons et al., 2015). Full scale BMA is both complicated and time-consuming which led Buckland and colleagues (1997) to propose simpler model averaging methods that also have been used within the BMD field (Faes et al., 2007; Moon et al., 2005; Piegorsch et al., 2013; Wheeler and Bailer, 2007; Wheeler and Bailer, 2009b). These frequentist procedures commonly relies on estimating model weights based on measures of the model fit such as the AIC, the corrected AIC (AICn), Bayesian Information Criterion (BIC) or the Kullback Information Criterion (KIC).

Various non-parametric and semi-parametric procedures for calculating the BMD has been suggested (Bhattacharya and Lin, 2010; Guha et al., 2013; Piegorsch et al., 2012; Wheeler and Bailer, 2012), but these have so far been used only rarely and not yet incorporated in regulatory guidelines.

The extent to which a model selection or model averaging procedure is conservative or anti-conservative can be estimated by calculating the coverage rate for the BMDL, i.e. how often the BMDL is lower than the true BMD, by using Monte Carlo simulations. Theoretically, the coverage rates should be 95%, but it can be substantially different if the model selected is different from the "true" model. West and colleagues (2012) have shown that relying solely on the AIC for modeling quantal data can lead to substantial undercoverage. Below expected coverage rates have also been noted in passing in a study on continuous data (Slob et al., 2005).

### 1.1.4 Experimental designs

The efficiency of the design of a toxicological experiment can be evaluated either by Monte Carlo simulations (Kavlock et al., 1996; Shao and Small, 2012; Slob, 2014b; Slob et al., 2005; Weller et al., 1995) or by evaluating or minimizing a design criterion such as the expected variance of the parameters (Dette et al., 2009; Holland-Letz and Kopp-Schneider, 2015; Krewski et al., 2002; Kuljus et al., 2006; Weller et al., 1995). Published reports involving any of these approaches are summarized Table 3 in the Results section. Öberg (2010) suggested that animal distress could be taken into account when investigating study designs, this has not yet been done.

## 1.2 ANIMAL EXPERIMENTS

Animal based research, carried out since the time of ancient Greece (Hajar, 2011), has been criticized, not least by the anti-vivisectionist movement that started in Britain during the nineteenth century (Rollin, 2006) and off-shoot organizations.

Today, animals are widely used in medical research and safety testing of chemical. It has been estimated that more than 100 million experimental animals according to the EEC definition (EEC, 1986) of animals and experiments, were used worldwide during 2005 (Taylor et al., 2008). While some animals experience little or no distress and/or pain in this context, others are subjected to significant discomfort. During 2013 in Canada 38.2% of the laboratory animals used were reported to be subjected to procedures that could potentially cause moderate-to-severe distress or discomfort, such as major surgical procedures under general anesthesia with subsequent recovery or exposure to drugs and chemicals at levels that impair physiological systems. 2.5% were reported to be subjected to procedures that could potentially cause severe pain near, at, or above the pain tolerance threshold for conscious animals, such as  exposure to drugs or chemicals at levels that (may) markedly impair physiological systems and which cause death, severe pain, or extreme distress (Canadian Council on Animal Care, 1991; Canadian Council on Animal Care, 2015). Although this categorization is based on a precautionary approach and the percentages are therefore likely to be overestimations, they are nonetheless disturbing.

### 1.2.1 Legislation concerning animal experiments

The laws regulating usage of animals in research are mostly regulated with a utilitarian perspective where the chance of positive outcome is weighted against the risk of harm (Vieira de Castro and Olsson, 2015). In the USA researchers must adhere to the United States Animal Welfare Act (USAWA) (US Department of Agriculture, 2013), which does not however, cover rats or mice bred for research purposes, that constitute the lion's share of all animals used in research. Within the EU, researchers must follow the Directive on Animals used for scientific purposes (EU, 2010) and all animal experiments must be impartially pre-evaluated to ensure that the benefits of the experiments outweigh the harm to the animals. In

Sweden this evaluation is performed by one of the six regional Animal Ethics Committees (AECs), consisting of both researchers and laypersons including representatives for animal welfare organizations.

Other legislation influences the utilization of experimental animals. The EU's regulation of chemicals, REACH, demands the identification and management of the risks linked to all chemicals imported into, or produced in the EU in an annual quantity larger than one ton (EC, 2006a) and Rovida and Hartung (2009) have estimated that fulfilling these demands could potentially require the use of 54 million animals before 2018. The REACH legislation do however encourage the use of animal free methods (EC, 2006b) and animal-free risk assessment, such as read-across, are being utilized to a greater extent (ECHA, 2014; Spielmann et al., 2011)

### 1.2.2  3R

The principle of the 3Rs (replacement, reduction and refinement) for animal experiments was launched in 1959 (Russell and Burch, 1959) and has now been incorporated in governmental legislation in the EU (EU, 2010).  The definition of the 3Rs has changed since the original definitions by Russel and Burch. The 3R Declaration of Bologna, defined:

"Reduction alternatives are methods for obtaining comparable levels of information from the use of fewer animals in scientific procedures, or for obtaining more information from the same number of animals.

Refinement alternatives as methods which alleviate or minimize potential pain, suffering and distress, and which enhance animal well-being.

Replacement alternatives as methods which permit a given purpose to be achieved without conducting experiments or other scientific procedures on animals." (Executive Committee of the Congress, 2000)

In many cases the 3Rs are positively correlated. For instance, refinement methods leading to less animal distress often also result in less variable data so that fewer animals are required to achieve acceptable scientific power (reduction). However, sometimes the 3Rs can correlate negatively, e.g. surgical implantation of telemetry devices for continuous monitoring, which may cause distress, can produce better data and consequently reduce the need for animals. The conflict between refinement and reduction is also evident in connection with toxicity testing. Lowering doses represents a refinement, but then the statistical power of the test will also be lowered so that more animals are required. In such situations, there is a lack of guidance concerning which R to prioritize (de Boo et al., 2005).

For an example, the EU directive on the protection of animals used for scientific purposes states that the benefits expected must outweigh the ethical cost i.e. a cost-benefit analysis must be performed for each individual animal (EU, 2010). However, this directive provides no general guidance concerning the relative priorities that should be assigned to reduction and

refinement. Consider a hypothetical example: A planned experiment that could be quite distressful involves two individual rats named Sprague and Dawley, among others. In experimental setup 1 both would be among the animals used and the expected benefit is considered to be larger than the ethical cost. If either Sprague or Dawley were removed, scientific power would be lost and no conclusion could then be drawn from the experiment. Therefore, if the cost-benefit analysis for the entire experiment is acceptable, it must include both Sprague and Dawley and their use is thus also acceptable on the basis of individual cost-benefit analysis.

The alternative experimental setup 2, offers the same potential benefits as setup 1, but Dawley no longer needs to be used. However an additional blood sample has to be taken from Sprague, thereby increasing his distress slightly. This experiment would still be considered ethically acceptable if the overall stress experienced by Sprague is outweighed by the benefits, but which setup should be chosen? The EU directive states that the individual animal must be considered, but Sprague suffers more in setup 2 whereas Dawley suffers more in setup 1. Such a decision must be based on ethical, not legal considerations.

### 1.2.3 Ethical considerations regarding reduction versus refinement

The choice between reduction and refinement is not entirely straightforward, especially since experimental animals are usually killed at the end of the experiment. Moreover, these animals are bred for this specific purpose and in the long run a reduction in the usage of animals will lead to fewer animals being born. It therefore becomes a conflict between quantity and quality of life, as discussed further by Sandøe and Christiansen (2007) .

Some argument support prioritization of refinement over reduction under all circumstances. For instance, animal rights philosopher Tom Regan advocates the worse-off principle in general whenever there is a conflict between rights. This principle states that if we must choose between overriding the rights of many or a few, it is better to override the rights of many, if harming the few will leave them worse off than the any of the many would be if the other option was chosen (Regan, 1983).

In many instances a hedonistic utilitarian, who are interested in the maximizing the amount of pleasure or wellbeing, have good reasons to prioritize refinement over reduction. If the experimental animals have a life generally worth living, reduction is even ethically questionable since it will lead to fewer animals being born and therefore less total wellbeing, a conclusion that most people probably find counterintuitive. Furthermore, it is not clear whether it is generally preferable to have more individuals, even if each has a positive wellbeing.  One problem with maximizing total happiness in this manner is that it easily leads to the so-called Repugnant Conclusion (RC) as follows:

In Figure 4, the heights of the bars represent the positive wellbeing of a group of individuals, and the width represents the number of individuals in each group. In Scenario A only contains individuals with a high wellbeing. Scenario A+ is similar to scenario A, but with the addition of more individuals with less, although still positive wellbeing, so in scenario A+ the

total wellbeing is greater. Scenario B contains the same number of individuals as A+ each with the same wellbeing which is slightly higher than the average in A+, it is then reasonable to consider B preferable to A+. If B is better than A+ and A+ is better than A, then B should be better than A.

If this process is repeated almost indefinitely, scenario Z in which an enormous number of individuals have a wellbeing that is barely high enough to be better than not living, will be reached. Z, which involves the most wellbeing, should be the best scenario of them all. This conclusion, that a world in which everyone has a life barely worth living, is better than a one which smaller number of individuals have a wonderful life is referred to as the Repugnant Conclusion (Figure 4).

Those who defend utilitarian welfare aggregation have dealt with this conclusion either by arguing that there is something wrong in the arguments involved (Blackorby et al., 1997; Ng, 1989) or that it is not actually repugnant (Ng, 1989; Tännsjö, 2002). Carlson (1998) has discussed a similar situation in which all cases (A-Z) include negative wellbeing with lives worth avoiding and concluded that those who are worse off should reasonably be given higher weight when aggregating the wellbeing. In an animal experiment this would mean that refinement should generally be given higher priority than reduction, although not in all cases. There will be situations where it is better for fewer animals to experience more negative wellbeing than for more animals to experience less negative wellbeing.
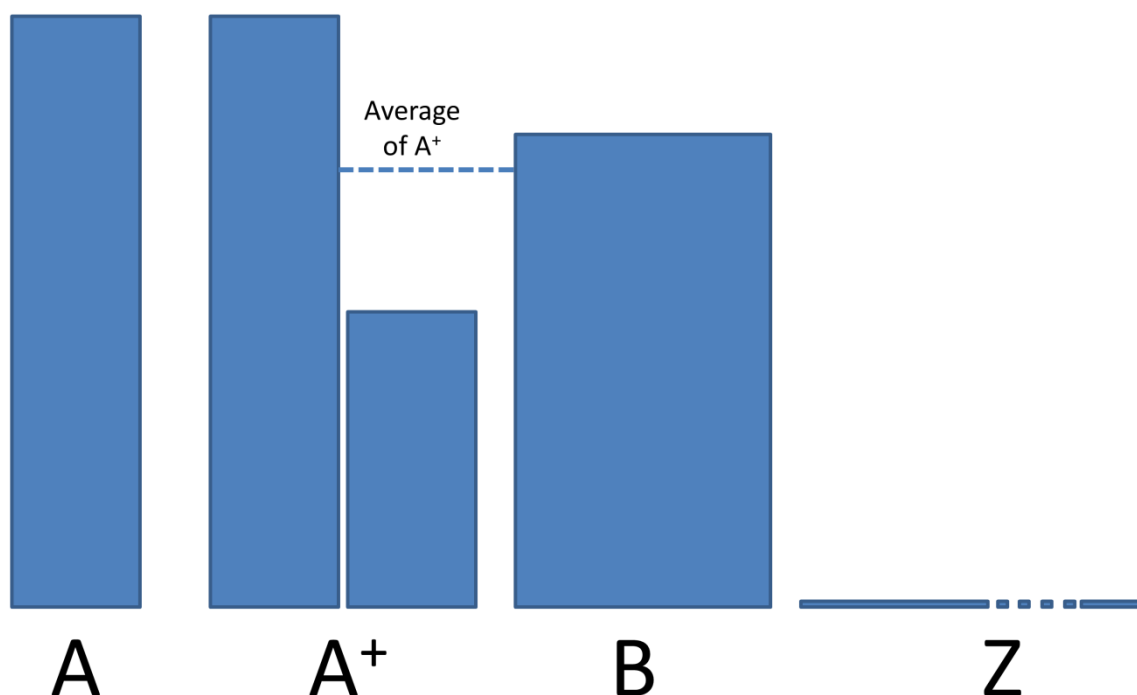


**Figure 4. The Repugnant Conclusion. See text for explanation.**

12

If killing animals always should be avoided it can be considered that reduction always should have precedence over refinement. This "badness-of-killing argument" has however been critized as it is not in line with how animal ethics are handled in other parts of society, such as concerning production of meat (Hansen et al., 1999; Olsson et al., 2012).

In the absence of normative prioritizations concerning reduction and refinement, ethical decisions are made on case by case basis, which implicitly or explicitly requires a measure of ethical cost, preferably on a cardinal scale where the ethical cost of using an animal is proportional to the ethical severity of that use. This scale must also be additive so that the ethical cost of using a number of animals is equal to the sum of the ethical cost for each individual animal.

There are few point and score systems for the quantitative ethical evaluation of animal experiments described to date (Porter, 1992; Stafleu et al., 1999) but their cardinality is questionable at best (Table 2). Moreover, their primary aim was to guide harm-benefit analysis rather than specific choices between reduction and refinement.

**Table 2. Ethical cost points concerning the number of animals in the scoring system developed by Porter (Porter, 1992). This scoring is not cardinal since the cost is not proportional to the number of animals.**

| Animals used | Cost points |
| --- | --- |
| 1-5 | 1 |
| 6-10 | 2 |
| 11-20 | 3 |
| 21-100 | 4 |
| >101 | 5 |

Other, more detailed scoring sheets for clinical signs of pain, distress and suffering have been developed for use with laboratory animals (Morton and Griffiths, 1985; Scharmann, 1999). These are often semi-quantitative and employed to determine when the animals should be given analgesia or euthanized because their suffering exceeds what is deemed acceptable. Accordingly, these scoring schemes involve the assessment of individual animals during the course of an experiment and they cannot in their present form be used directly to decide between reduction and refinement in connection with experimental design.

A set of cardinal weights for animal experiments could be developed using an approach similar to the Person Trade-Off (PTO) technique, originally known as the equivalence technique, that has been used to derive Quality Adjusted Life Years (Murray and Lopez, 1996). With the PTO interviewees are asked questions such as: "If there are x people in

adverse health situation A and y people in adverse health state B and if you can only help (cure) one group (for example due to limited time or limited resources) which one would you choose to help?"(Torrance, 1986).

# 2 AIMS

The general objective of this thesis was to improve BMD modeling in relationship to the use of experimental animals in toxicological experiments. The specific objectives were as follows:

**Paper I**

To investigate under what circumstances and how frequently BMDL coverage with continuous data are below the nominal level. A secondary aim was to investigate the coverage of the NOAEL in relation to the true BMD.

**Paper II**

To investigate whether and under what circumstances experimental designs involving dose groups of unequal size can both result in less animal distress and provide more reliable estimates of the BMD for quantal data.

**Paper III**

To investigate whether and under what circumstances experimental designs involving dose groups of unequal size can both result in less animal distress and provide more reliable estimates of the BMD for continuous data.

**Paper IV**

To determine cardinal ethical weights for toxicity testing, as well as investigate how these differ between different categories of responders.

**Paper V**

To investigate the impact of taking ethical cost into consideration in connection with optimization of dose-response studies.

# 3 METHODS

In **Paper I** the BMDL coverage rates were investigated by simulating continuous data from a sigmoidal exponential curve and then calculating the BMDL values with a standard nested set of exponential models (Figure 3). One of these models was selected on the basis of likelihood-ratio tests as described by Slob (2002). Utilizing a model from earlier simulations studies as the true underlying model, Monte Carlo simulations were performed to investigate how often the BMDL and the NOAEL was higher than the true BMD, and how much higher they were

In **Paper II** the effects of using unequal numbers of animals in the different dose groups for toxicological studies, with in total 200 animals, on the quality of BMD estimates were investigated using Monte Carlo simulations on quantal data. Six different dose-response models commonly used in BMD calculations for quantal data were used in the simulations and in the re-estimations. All six were used as true models, by fitting them to two different datasets from an NTP (National Toxicology Program) cancer study on furan (NTP, 1993); one on the incidence of hepatocellular carcinoma, without background incidence, and one on the incidence of mononuclear cell leukemia, where there was a background incidence of 16%. Thus, 12 (2×6) different "true" curves were used in the simulation. Nine different dose placements were evaluated, ranging from very low doses to very high doses, each with 85 different distributions of animals between four dose groups. An AIC-based model averaging approach was used and the performance of a specific design evaluated using the root mean squared error (RMSE) of the AIC-averaged BMD estimates and by calculating the ratio between AIC-averaged BMDU and the AIC-averaged BMDL. The animal distress was assumed to be proportional to dose.

In **Paper III** the effects of using unequal numbers of animals in the different dose groups for toxicological studies on the quality of BMD estimates were investigated using Monte Carlo simulations on continuous data. The simulation step was based on four different hypothetical "true" curves. The curves included two sigmoidal curves, one that either clearly levelled off within the covered dose-effect span and one that barely leveled off within the dose-effect span as well as two models that did not level off. Designs with either 40, 80 or 200 animals in total were evaluated as these are common study sizes in OECD guidelines (OECD, 2012). Nine different dose placements were evaluated, ranging from very low doses to very high doses, each with 85 different distributions of animals between four dose groups. An AIC-based model averaging approach was used and the performance of a specific design evaluated using the root mean squared error (RMSE) of the AIC-averaged BMD estimates. The animal distress was assumed to be proportional to dose.

In **Paper IV** members of the Swedish Animal Ethics Committees (AEC) were interviewed via telephone concerning how they prioritized reduction versus refinement in connection with toxicological experiments. The interviews were based on a fictitious one-week study in rats. The committee members were asked to evaluate the ethical impact of nine different clinical signs, each having one mild and one severe variant, and for each sign an ethical weight was

determined as how many animals free from clinical signs would entail the same ethical cost as a single animal experiencing the sign. The ethical weights assigned by the different member categories (researchers, political representatives and laypersons representing animal welfare organizations) were evaluated with Kruskal-Wallis tests.

In **Paper V** various study designs were evaluated using Monte Carlo simulations of quantal data. The designs did not have the same number of animals as is the case with most studies on experimental design, instead they had the same estimated ethical cost. The incidence of hepatocellular carcinoma and mononuclear cell leukemia in male rats (i.e. the same data as in **Paper II**), were used to define the true curve. The loglogistic model was included as a "true" dose-response model in the simulation step. All six models used in Paper II were fitted to the simulated data, and the BMD estimated from each model were averaged using an AIC based methodology. Several different study designs were evaluated, all having approximately the same ethical cost as a study with 200 animals evenly distributed to four dose groups at a medium-low dose placement. The ethical cost of a design was estimated in the basis of with ethical weights of 1, 4, 16, 64 or 256, based on the result from paper IV. The "true" curves were used to define the ethical cost of the studies, i.e. this cost of exposing to a certain dose depended on the response at that dose. Both datasets on hepatocellular carcinoma, without a background incidence, as well as on mononuclear cell leukemia, with a background incidence, were used to define the true dose-response curve as well as to define a dose-ethical cost-curve.

# 4  RESULTS AND DISCUSSION

## 4.1  MODELS FOR CONTINUOUS ENDPOINTS

In **Paper I** we found that coverage rates (i.e. how often the estimated BMDL is lower than the true BMD) depended on the dose placements and the assumed coefficient of variation (CV). These rates were in many cases considerably lower than the theoretical 95%, indeed in some scenarios as low as 20%, due to the exclusion of a ceiling parameter. With lower doses, the ceiling parameter was often excluded from the model selected, often resulting in coverage rates that were quite poor. On the other hand the BMDL values were on only slightly higher than the true BMD. With higher doses, the ceiling parameter was identified more often, so the coverage rates were closer to the expected 95%. However, when the ceiling parameter was excluded, the BMDLs were even less protective. Although the coverage of the BMDLs were somewhat disturbing, the situation with the NOAELs was generally worse.

Less than nominal coverage rates are especially problematic from the point-of-view of the EFSA which generally supports a conservative approach to model selection, advocating the model with the lowest BMDL in cases where several model fits the data (EFSA, 2009). Thus the BMDLs selected using EFSAs guidelines would be expected to be conservative in general, but this is not the case as they don't advocate the use of the model with lowest BMDL within a nested set, but instead advocating the selection one model from a nested set using likelihood-ratio tests.

Our findings on the BMDL coverage rates favor the use of a ceiling parameter (c) in dose–effect analysis, although this is not always entirely unproblematic. Within the nested set of exponential models, the ceiling parameter will be excluded if the data do not provide a significant amount of information about this parameter. Obviously, if the c-parameter is always included, the likelihood curve will sometimes be very flat due to the lack of information about the parameter which will lead to problems with model convergence.

However, a lack of convergence should not automatically lead to dismissal of the result of the model fit, since it is the exact value of the c-parameter that is not of interest in BMD modeling, but rather the BMDL which is not much influenced by the exact value of an uninformative c-parameter.

More problematic is the fact that the overparametrization increases the risk of errors in the confidence interval estimated by the profile likelihood procedure, if the shape-determining d-parameter is included as well.  The risk of obtaining such erroneous confidence intervals could potentially be avoided by starting the numerical estimations during the confidence interval estimations at different points in the parameter space, at least in those cases where there is a sharp drop in the profile likelihood curve.

Another issue with the 4-parameter models, that includes both a ceiling parameter (c) and a shape-determining parameter (d), is that the confidence intervals can be very broad, especially if the d-parameter is unrestricted. From a precautionary perspective such a wide confidence interval and very low BMDL could be reasonable. The BMDL should reflect the lower limit of the risk and if the data do not support the absence of effects at low doses, then the BMDL value should be low. However, from a regulatory perspective it would be unfortunate if RfD based on continuous data differ from RfD obtained with quantal data (Crump, 2002), and since the 4-parameter models are much more flexible RfD based on these could be considerably more conservative.

Broad confidence intervals could be restricted by including prior information concerning the dose-effect relationship. This could either be done by including historical datasets on the same endpoint (Slob, 2014a; Slob and Setzer, 2014) or by including endpoint specific limits to the model parameters. However, both of these solutions require use of prior information and increases the demand of expert knowledge needed for the BMD analysis.

It is also possible that always including d-parameter and allowing it to be lower than 1, instead of always including the c-parameter, may provide a model with enough flexibility to result in reasonable coverage rates.

Yet, another approach would be to determine BMDs utilizing non-parametric approaches, but these can also result in unnecessarily wide confidence intervals (Slob and Setzer, 2014).

## 4.2   BMD AND EXPERIMENTAL DESIGN

Our major findings concerning study design in **Paper II** and **Paper III,** as well as relevant findings by others, are listed in Table 3. **Paper II**, with quantal data, indicates that it is important to include doses close to the targeted BMD, or a bit above the targeted BMD if there is a high background incidence. This is in line with the conclusions by Slob (2005)  and Kavlock and colleagues (1996). Shao and Small (2012) found that the best design for one of their two datasets had the lowest dose group much higher on the dose-response scale. However, Shao and Small used a different quality metric (the difference between the 95th and 5th percentile of the BMD estimates) and that can possibly contribute to the difference in the results.

Moreover, **Paper II** and **Paper III** indicate that it is important to include higher doses with a clear response as well, which is in agreement with the findings of others (Dette et al., 2009; Holland-Letz and Kopp-Schneider, 2015; Krewski et al., 2002; Shao and Small, 2012; Slob et al., 2005).

Our observation that dose groups of unequal size, with a larger number of animals close to or above the BMD, is generally in agreement with previous reports (Dette et al., 2009; Kavlock et al., 1996; Krewski et al., 2002; Weller et al., 1995).  However, **Paper II** indicates that this scientific gain is quite small with quantal data, as was also observed by Shao and Small (2012).

The other studies did not state how big the difference was between an equal distribution of animals and an optimal uneven distribution. **Paper III** showed that with continuous data this gain was greater and moreover, it seems as the control groups is more important for continuous data than for quantal data.

Using distress as a criterion for evaluating designs has not been done before. In many of the scenarios in **Papers II** and **III,** the best design with more animals receiving a dose close to the BMD, showed less estimated distress. It was shown that although the improvement was quite limited there is a potential to use BMD-aligned experimental design as a means to refine toxicity testing. However, this was most clearly evident when the doses were high in general, in the range where a clear effect could be observed and where distress also can be expected to be higher.

Some of the previous studies have included parameter or model uncertainties in the modeling. Shao and Small (2012) combined two models as "true" models, but did not take into account potential uncertainties in the location of the dose-response curve. Kuljus and colleagues (2006) as well as Holland-Letz and Kopp-Schneider (2015) included uncertainty in the steepness paramater, but their suggested optimal designs did not take into account the uncertainty in the dose placement. Dette and colleagues (2009) took uncertainties into account when determining optimal designs in a part of their study. However the uncertainties were very small, they are much larger when designing real toxicity experiments. In **Paper II** and **Paper III** we to some extent took dose placement uncertainty into account by investigating the same designs at different dose placements, although we did not consider a continuous range as Dette and colleagues did. Further investigation of the designs of studies should ideally take into account realistic uncertainties in the parameter estimates, and ideally also model uncertainties.

Various approaches on how to incorporate prior information or how to combine different datasets have been described (Slob and Setzer, 2014; Wheeler and Bailer, 2009a). The influence on such approaches on experimental design aligned to BMD has however not been examined and warrants investigation. Wang and colleagues (2013) proposed optimization of the design of a sequential experiment by performing by designing sequential design, i.e. the second stage is designed after the first, but this approach does not take variability between the experimental stages into account.

This issue of differences between experimental stages performed in the same laboratory is analogous to designing new studies on the basis of previous ones. The uncertainties are larger in the latter case, but not fundamentally different. Since these uncertainties need to be included in the development of experimental designs their quantification in desirable.

**Table 3. Summary of previous and present studies concerning the estimation of BMD. "-" = Not investigated.**

| Article | Methods | | | | | Results | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Type of data | Evaluated simulated data or minimized expected variance? | Was model uncertainty included the estimations ? | Were BMDLs calculated ? | Was animal distress considered ? | Did more dose groups give better estimates? | Are the estimates better with a dose closer to the BMD? | Are high doses important? | Does unequal distribution of animals have an impact on quality? |
| (Weller et al., 1995) | Quantal | Both. | No | Yes | No | More groups resulted in better accuracy, but worse precision. | - | - | Yes, few animals needed at the high dose. |
| (Kavlock et al., 1996) | Quantal | Evaluated simulated data | No | Yes | No | Not necessarily, it depended on the situation. | Yes | Doses close to the BMD are more important. | - |
| (Krewski et al., 2002) | Quantal | Minimized expected variance | No | No | No | 3-4 groups are more efficient than 5-7 groups. | Sometimes, but not always. | Yes | Yes, fewer animals needed at the high dose. |
| (Dette et al., 2009) | Quantal | Minimized expected variance | No | No | No | Sometimes 4 doses are better than 3. | - | Yes | Yes, few animals are needed at the highest dose. |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **(Shao and Small, 2012)** | Quantal | Evaluated simulated data | Yes, used two models weighed with a BMA approach. | No | No | Yes. | On one occasion the best design involved a dose almost as high as $ED_{50}$. | Yes | Only minor effect. |
| **(Slob, 2014b)** | Quantal | Evaluated simulated data | No | Yes | No | Not clearly. | - | - | - |
| **(Slob et al., 2005)** | Continuous | Evaluated simulated data | Yes, model selection based nested set of models | Yes | No | More dose groups reduced the risk of poor dose placement. | Most often, yes. | Yes | - |
| **(Kuljus et al., 2006)** | Continuous | Minimized expected variance | No | No | No | > 4 dose groups better when parameter values were uncertain. | No, at least not a strong trend. | - | - |
| **(Holland-Letz and Kopp-Schneider, 2015)** | Continuous | Minimized expected variance | Yes | No | No | Yes, > 4 dose groups were preferred when parameter values were uncertain. | No | Yes | Yes, but there was no clear trends regarding how to distribute the animals. |

| | | | | Included in this thesis | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Kalantari et al (Paper II)** | Quantal | Evaluated simulated data | Yes, 6 models weighted together with an AIC-based model averaging approach. | Yes | Yes | - | Yes | Yes | Small effect on BMD, but potential ethical benefit |
| **Ringblom et al. (Paper III)** | Continuous | Evaluated simulated data | Yes, 4 models together with AIC-based model averaging approach. | No | Yes | - | Not necessarily, but it was advantageous with many animals closer to the BMD. | Yes | More pronounced effect on BMD than in Paper II. |
| **Ringblom et al. (Paper V)** | Quantal | Evaluated simulated data | Yes, 6 models together with an AIC-based model averaging approach. | Yes | Yes | - | Yes | Yes, if there was a background incidence of distress. Less so without a background incidence of distress. | Yes, large effect since few high dose animals was converted to more low dose animals. |

## 4.3 REDUCTION VERSUS REFINEMENT

In 47 interviews with members of the Swedish AECs in **Paper IV**, priorization between reduction and refinement varied widely. One researcher and one political representative always prioritized reduction and 3 researchers, 1 political representative and 2 representatives of animal welfare organizations always prioritized refinement. The responses of the remaining 39 participants implied that a limited increase in animal numbers in some cases could be acceptable if the individual animal distress was reduced.

The median ethical weights, that is how many animals not showing a clinical sign that entailed the same ethical cost as 1 animal with the clinical sign, was 2-4 for the milder version of the signs and 5-20 for the more severe version of the signs. There were no statistically significant difference between the magnitudes of the ethical weights assigned by different member categories of members (researchers, political representatives and representatives of animal welfare organizations) the within group variation was large compared to the between group variation. There where however a small trend that the political representatives assigned lower ethical weights than the other committee members.

These similarities between the groups raise the question as to whether there is any reason to include laypersons in the AECs. Personally, I believe, in agreement with others (Hansen, 2013), that the laypersons play an important role and that the committees should not consist of researchers alone. When evaluating research protocols, the members must weight the harm of an animal experiment against the scientific benefits. Our participants only weighted harm against harm, which is not the same thing.

8 participants (5 researchers, 1 politically nominated layperson and 2 laypersons nominated by animal welfare organizations), found the questions to be too hard to answer and did not complete the interview, as also happens in connection with PTO studies on human health (Damschroder et al., 2007). Such studies involve making decisions concerning the health of other humans and it is not surprising that many find PTOs difficult and unpleasant (Nord, 1995), nor is it surprising that this is the case for questions regarding animal experimentation as well, which is potentially even more sensitive than questions regarding health care.

In an examination, on the balance between reduction and refinement, Franco and Olsson (2014) asked participants in a Laboratory Animal Science course if they ethically preferred performing a stressful experiment with no permanent effects 20 times on one animal or once in 20 animals. If the animals were mice, a slight majority preferred refinement, using more animals, whereas if the animals were primates or dogs more favored reduction. Franco and Olsson note that this difference might reflect considerations other than purely ethical ones, such as financial and logistical considerations, but some ethical differences might still be truly ethical. This indicates that the size of ethical weights could be species specific. Also, completely different clinical signs may be needed for different species.

It is also quite possible that cultural differences between countries regarding reduction versus refinement exist. For instance, there are considerable differences among the residents in the different countries of Europe concerning the opinions on human euthanasia (Cohen et al., 2006). This question has, of course, other dimensions than animal euthanasia and reduction versus refinement in animal experiments, but the value of a life and wrongness of killing is involved in both cases. According to the report by Cohen, Swedes are more positive to euthanasia than the inhabitants in most other European countries, and one wonders whether this might also be the case with respect to refinement of animal experimentation. If so, the ethical weights determined in our study would be expected to be higher than if the study was performed in a similar test population in a different country, for instance in southern Europe.

Previously statements on reduction and refinement have mostly been of qualitative nature. However in some situations qualitative statements are not informative enough, for instance when it comes to evaluating several experimental setups or test strategies. The determined ethical weights are up for criticism, for instance regarding their accuracy. However being up to criticism in some sense positive, compared to mere qualitative statements that often lack specificity and evaluability.

In **Paper V,** we used the quantitative ethical weights determined in **Paper IV** to investigate how toxicity tests can be designed taking into account both the number of animals used and the experiences of the individual animals. We evaluated ethical weights of 1,4,16,64 and 256. The results show that the optimal dose placement was heavily dependent on the ethical weight of the sign determining the ethical cost and the background incidence of that clinical sign. When the distress of the individual animal was not considered at all (ethical weight=1) it was preferable to place the doses relatively high on the dose-response scale with the mid dose group around or even above the $ED_{50}$. Already the use of an ethical weight of 4, made it generally preferable to have the mid dose placed below the true BMD ($ED_{10}$), if there was no background incidence of distress and no background incidence of the toxicological endpoint. However, if it was assumed that there was distress present already in the control group, there is not a lot to gain by moving animals to lower doses.

The use of even higher ethical made it more advantageous to use more animals at lower doses, but it was only in one case where it was preferable to use the lowest dose placement tested, with the high dose group around the true BMD. In that case there were more than 1000 animals in the study and it seems unlikely that such study would be performed in practice.

The ethical weights determined in **Paper IV** were based on one-week experiments on eight week old rats, while the dose-response data in **Paper V** was from a two-year study. It is not obvious that these weights are directly transferable in this manner. In a two-year study the animals could suffer distress for a longer period of time, but they also live longer, i.e. might have longer periods of life worth living as well, balancing the enhanced stress out from a utilitarian perspective.

The degree to which the ethical weights would need to be adjusted to be appropriate for a 2-year study, remains an open question. On the other hand we tested a wide range of ethical weights (1,4,16,64,256) in **Paper V,** so even if the weights are changed, it is still possible to draw conclusions about study designs.

Are the results presented in **Paper V** relevant also for shorter studies, even though the dose-response data are from a longer study? In principle, they should be. The background responses can surely be the same for endpoints relevant to shorter exposures. At the same time the slope of the dose-response curves could differ for different endpoints and different experimental setups. Nothing has yet been published concerning the shape of the quantal dose-response data, although Slob and Setzer mention that they are working on this (Slob and Setzer, 2014). In the absence of such data, I find no compelling reason to believe that the shape of the dose-response curve for carcinogenicity should differ markedly at higher doses compared to other quantal responses.

## 4.4 METHODOLOGICAL CONSIDERATIONS

### 4.4.1 Assumptions regarding dose-response models

In **Papers I-III** and **V** we studied BMD modeling using Monte Carlo simulations meant to emulate real life experiments. Investigating the performance of different designs by performing the experiments with real animals may look favorable for obvious reasons. However, computer simulations offer the advantage "true" dose-response relationship is known by definition, providing a reference for the results. Also, with computer simulations it is possible to investigate many possible situations (i.e., shapes of dose-response curves, combinations of designs, etc.) compared to what can be realistically evaluated with real toxicity data. The results obtained with a simulation approach are of course dependent on the assumptions made in the simulations, e.g. the true models used, the values of the parameters and variation in these models etc. In these papers of this thesis we employed several different assumptions depending on the specific aims of each project.

In **Papers II** and **V** the true models originate from two actual datasets with different background incidences in an NTP cancer study on F344 rats exposed to Furan (NTP, 1993). In **Paper II** six different models of varying steepness were fitted to each dataset, giving rise to 12 different dose-response curves, with different steepness. The F344 strain is an inbred strain and it is therefore likely that the dose-response curve is steeper than it would be from an outbred strain or human population. In **Paper V** we fitted only the loglogistic model to the two different endpoints, giving rise to two different true curves, one without and the other with a background incidence.

In **Paper I** we employed a dose-effects model already used in a similar simulation study (Slob et al., 2005), the choice of parameters and CV in previously published article was based on a database of dose-effect data. We used CVs of 5%, 10% and 15%. It has been suggested that the size of the CV covaries with the difference between the maximum effect and the background effect (Slob, 2014b) . If so, it is likely that 5% scenario is the most realistic one.

In **Paper III**, we used four hypothetical curves (exponential, Hill, power and polynomial). The exponential and Hill models are realistic according to Slob & Setzer (2014). We also used the power and polynomial models in both the simulation and estimation step. It could have been argued that both of these models should have been omitted on the basis of the findings by Slob and Setzer (2014), who showed that a vast array of dose-effect relationships can be adequately described using the four-parameter exponential and Hill models. The power and polynomial models were included anyway since it is common practice to use these in dose-effect modeling.

While the results in **Papers I** and **III** are dependent on the model parameters, it can be demonstrated that the results from the simulations based on continuous data are valid also for situations with different CVs, as long as the CES and ceiling parameter (c) is changed appropriately as well (see Slob, 2005 for details).

In contrast to the case for quantal data, the choice of CV and the assumptions regarding the distribution of the data both matter for continuous data. We assumed that the data are distributed lognormally. Shao and colleagues (2013) demonstrated that the assumption regarding normality or lognormality has limited impact when the CV is small (CV=10%), which was the case in **Paper I** and **Paper III.**

## 4.4.2  What constitutes a good design?

Different investigations in the literature have used difference methods and different quality metrics to evaluate different experimental designs. In **Paper II, III** and **V** we explored experimental designs using Monte Carlo simulations and we employed the RMSE as the primary quality metric. Minimizing a design criterion, such as the expected variance of the parameter estimates, was an alternative to the simulations, but simulations were chosen as they reflect actual experiments more closely. Minimizing the expected variance gives no information regarding the frequency of statistically significant dose-response relationships or the BMDLs.

We employed RMSE of the BMD as the primary quality metric in our simulations. The advantage of using the RMSE is that it measures both accuracy and precision of the BMD estimate. The RMSE (or MSE) has also been used as a quality metric in earlier simulation studies involving BMD estimations (Fung et al., 1998; Guha et al., 2013; Kavlock et al., 1996).  It might appear to be more realistic to primarily employ a metric based on the BMDL rather than the BMD since the lower confidence interval is the value used in risk assessment and BMDU/BMDL or BMD/BMDL ratios are commonly used to assess the precision in BMD analysis. However, these metrics only assess the apparent precision and not real precision and they should therefore be used with caution as quality metrics in simulations. For instance the BMDU/BMDL ratio can be very low, indicating good estimation, while the "true" BMD is actually outside of the BMDL-BMDU interval (as shown in **Paper I**).

An alternative approach has been proposed by Slob(2014a). He proposes that minimizing the "true"BMD/BMDL ratio provide good evaluations in BMD simulations. Following this suggestion literally is, however, not recommended since it implies that higher BMDLs are always better and that anti-conservative approaches are always preferred. It would perhaps be sounder to use the coverage rates or the "true"BMD/BMDL$_{95th\ percentile}$ as quality metrics. Although a coverage of 95% and a "true"BMD/BMDL$_{95th\ percentile}$=1 seems favored it is unclear whether a coverage of 90% and a "true"BMD/BMDL$_{95th\ percentile}$=0.9 is better or worse than a coverage of 99% "true"BMD/BMDL$_{95th\ percentile}$=1.3. In addition, results based on coverage rates will be heavily dependent on the model selection or averaging procedure. The procedure we applied to weight BMDLs together is not formally correct as the BMDL is "not an independent random variable but a statistic of the variable BMD"(Shao and Gift, 2014), although it has been suggested and used by others as well (Bailer et al., 2005a; Bailer et al., 2005b; Wheeler and Bailer, 2009b). RMSEs based on BMD estimates are also influenced by the choice of selection or averaging procedure, but less so.

In **Paper III** some of the BMD estimates were very high due to the fact that the dose-effect curve was very flat. Consequently we set a limit on the BMDs for certain outlier datasets, treating them as exhibiting lower BMDs than they actually did. Otherwise, the quality of the designs would have depended solely on single outlier values. A different solution would be to keep all of the BMDs and calculate the Root Median Squared Errors instead of the more commonly used Root Mean Squared Errors, as medians are less sensitive to outliers.

In **Paper II** and **V** we left out models with poor fits in the model averaging, as suggested by Wheeler and Bailer (2009b). Accordingly for certain datasets no models gave an acceptable model fit and thus there were no BMDs or BMDLs for these datasets. In **Paper II** these simulations without a $BMD_{AIC}$ were excluded from the RMSE calculations. In **Paper V** we used an alternative approach where they were treated as having a $BMD_{AIC}$ to 10 interquartile ranges higher than the average $BMD_{AIC}$ for that particular design, in order to penalize simulations without a dose-response trend as false negatives are negative outcomes.

BMDU/BMDL ratios, a measure of apparent precision, were used as secondary quality metrics in **Papers II** and **V**, but not in **Paper III** where preliminary simulations indicated that confidence interval calculations sometimes resulted in erroneous BMDLs and BMDUs due to numerical problems when calculating the profile likelihood curve. With real data such erroneous confidence intervals can be identified by visual inspection of the profile likelihood curve, but in the present case it was not possible to visually inspect the output of all BMD calculations. In retrospect, we could have accepted the erroneous BMDL, since as long as they are relatively few in number, they distort the median BMDU/BMDL ratios only slightly.

### 4.4.3 Trade-off interviews

In **Paper IV** we determined cardinal weights for the ethical cost of animal experiments by trade off interviews with the members of the Swedish AECs. We chose the interview group based on that they are used to evaluate situations regarding animal ethics and they also include members of different backgrounds and beliefs. The participants in our study match the composition in the Swedish AECs fairly well when it comes to the fraction of members being researchers, politicians or animal welfare representatives and there were participants from all regional committees. It is still possible that those who agreed to participate are not representative. Furthermore, the representativeness of committee members may not be ideal. For instance, maybe a more ideal test population would include individuals with other backgrounds as well, such as ethicists and ethologists.

We conducted interviews, instead of written surveys, as trade-off questions are easy to misunderstand. For instance, in a PTO investigation involving written surveys by Ubel and colleagues (2002), two thirds of the responses of showed inconsistencies and had to be excluded. Consequently, interviews are the gold standard for trade-off investigations allowing the task to be explained more thoroughly and inconsistences to be addressed.

However, interviews can be quite time-consuming and moreover, involve a risk that the interviewer influences the participants (Damschroder et al., 2004). The answers obtained at

later stages of a PTO study can be influenced by anchoring these numerically to earlier answers (Ubel et al., 2001; Ubel et al., 2002). In other words once a participant has given a numerical value, the next choice will be anchored to that value. Alternatively, participants can anchor their values to numerical values provided, e.g. when a bidding game (ping-pong) methodology, a common search elicitation in trade-off studies is utilized. In such a bidding game the participants answers iterative yes/no questions, e.g. as follows:

*1.		Would 11 animals experiencing mild tremor entail a higher ethical cost than 10 animals experiencing severe tremor?*

*2.		Would 1 000 000 animals experiencing mild tremor entail a higher ethical cost than 10 animals experiencing severe tremor?*

*3.		Would 20 animals experiencing mild tremor entail a higher ethical cost than 10 animals experiencing severe tremor?*

*4.		Would 10 000 animals experiencing mild tremor entail a higher ethical cost than 10 animals experiencing severe tremor?*

and so on until a point of indifference in reached. Since our questions were not framed following the ping-pong approach, the participants could not have anchored their answer to a number provided by the interviewer. The ping-pong methodology is also more time consuming method. On the other hand the participants had to decide their point-of-indifference directly, which can be more difficult.

Possibly, our participants could have anchored their later answers to their earlier ones and we might have gotten different weights if the questions had been asked in a different order. Participant fatigue could have a similar effect. A solution to both of these problems would have been to ask the questions in random order, but that would have been stressing to the interviewer thereby increasing the risk for other mistakes and in addition it would have increased the documentation of the responses.

Others have shown that PTO responses often deviates from cardinal transitivity, i.e. that one such response cannot be accurately inferred from two other (Baron et al., 2001; Dolan and Tsuchiya, 2003; Schwarzinger et al., 2004; Ubel et al., 1996). We therefore included built-in checks for cardinal transitivity between the mild and severe clinical signs. There was, however, no direct check for cardinal transitivity between different types of signs, although many participants to some degree provided such checks explicitly by thinking out loud during the interview. Furthermore, it remains to be evaluated whether our questions exhibits a good test re-test reliability and whether the results are reproducible.

There might be a strong random element in the trade-off studies, but since random elements do not introduce bias, median equivalence numbers in large groups of people may be more reliable (Nord, 1995). Although our group was not so large, we interviewed at least a reasonable part (~24%) of the members of the Swedish AECs.

While the clinical signs were defined and described in the same manner to all participants they may nonetheless have been interpreted differently, which might have influenced the weights. This issue can potentially be solved by showing videos of animals experiencing the clinical signs, but many signs are difficult for an unskilled professional to interpret, so most laypersons in the AECs would likely have struggled interpreting such videos.

### 4.4.4 Other issues related to ethical cost of animal distress

An unavoidable weakness of our approach is that the ethical weights are assigned by humans, since we could not ask the animals about their opinion. Accordingly, the weights are not only subjective, but also assigned by subjects that are not ideal and whose appraisals could, for instance, be distorted by anthropomorphic tendencies. Such anthropomorphism could, for example lead to an ethical weight for "weight loss" that is too low since many humans would not mind losing a few kilos of weight themselves.

There are more objective physiological measures utilized to assess the stress experienced by animals during experiments, such as the grimace scales (Keating et al., 2012; Langford et al., 2010; Sotocinal et al., 2011) and non-invasive measurement of metabolites of stress hormone in feces and amylase levels in saliva (Kolbe et al., 2015; Matsuura et al., 2012). Alterations in stress hormone levels can however be caused by both pleasant and unpleasant situations (Dawkins, 2008). Moreover, no measure such as these could directly be used as ethical weights, since we cannot say anything about if having two animals with a certain facial expression or hormone level are equally regrettable as having one animal with a worse facial expression or higher hormonal level.

It is also possible to conduct preference tests concerning how much effort an animal is willing to put in to achieve something positive or avoid something negative. Such a study could be performed to evaluate the relative severity of some of the clinical signs. Such tests would, however, be ethically questionable, at best.

We based our ethical weights on clinical signs since these are recorded in toxicity tests for everyday assessment of animal welfare and determination of the suffering of the animal surpasses what deemed acceptable in the study, so that the animal should be humanely killed (OECD, 2002). Of course, other factors not picked up directly by our ethical weights, such as the size of the cages, presence of environmental enrichment and cage-mates also contribute to animal welfare (Balcombe, 2006). In addition, lack of clinical signs does not necessarily mean absence of distress. For example, animals can suppress the expressions of distress to deceive predators.

# 5  CONCLUSION

The current use of nested models in the determination of BMDs for continuous endpoints could lead to coverage rates below nominal level due to the fact that the simpler models with fewer parameters are not flexible enough. Since coverage rates below the nominal level leads to underestimation of the risk, models of lower order should be used with caution in risk assessment. In addition, it is clearly shown that the NOAEL approach is even more problematic.

To establish BMD values with high quality, it is important to include a dose located relatively high on the dose-response scale.  Employing dose groups of unequal size can also slightly increase the quality of BMD estimates or conversely allow the same quality with fewer animals. In general, it is preferable to place more animals in the dose groups around the true BMD, or a bit above the BMD if there is a high background incidence of the selected endpoint. Such designs could also be utilized to reduce the animal distress.

Prioritization between reduction and refinement, expressed as ethical weights for clinical signs, varies considerably among the member of the Swedish AECs. The median ethical weights were 2-4 for the mild versions of the clinical signs and 5-20 for the severe versions. Some participants assigned an ethical weight of 1 to all signs (always giving priority to reduction) while others assigned infinity to all signs (always giving priority to refinement). No statistically significant difference was observed between the three categories of committee members (researchers, political representatives and representatives of animal welfare organizations) regarding the magnitude of the ethical weights.

Ethical weights with cardinal properties can be used to explore designs for toxicity tests that optimize the ethical cost in terms of both number of animals and their distress. These optimized designs are heavily dependent on what constitutes the ethical cost, and the relative ethical importance of those costs. Using more animals, but at lower doses, can be ethically justifiable. Even though it can be ethically justifiable to use a very large number of animals at very low doses (all doses below the BMD), the large number of animals required render such an approach impractical in reality.

# 6  FUTURE RESEARCH PERSPECTIVES

Based on the results in this thesis several areas for further investigation and research have been identified. In general, the BMD approach and the underlying strategies for model selection need to be improved and harmonized. In addition, the alignment between BMD analysis and experimental design needs to be further studied and implemented in guidance documents. Moreover, the 3R-principles could be used as a factor when evaluating experimental design and approaches for dose-response modelling. The following paragraphs include specific suggestions and ideas for research studies within this field of research.

The best way to select a BMDL, to use as a PoD, from continuous data needs to be elucidated further.  This could be done by performing large studies that compare the effects of different approaches, similar to the ones performed in connection to their modeling averaging workshop (US EPA, 2015), including (e.g. non-parametric approaches, model averaging of the currently used models etc) on the coverage rates of the BMDLs.

Further investigations concerning how to design experiments on the basis of prior data, such as previous studies on similar compounds and the same endpoint are warranted.  Such investigations should ideally take into account parameter uncertainty, especially with regard to the potency parameter/dose placement. In this context, additional analysis of historical data as performed by Slob and Setzer (2014) would be valuable as would studies designed to quantify the uncertainties that can be expected when designing studies.

Our study on ethical weights in **Paper IV** is the first of its kind and there are numerous ways to expand upon it. First this investigation could be repeated in different settings, with participants of different types and/or from different countries. **Paper IV** also only considered a one week study in rats. The impact of other study durations and experiments concerning different species also needs to be further elucidated. Also, we focused on the clinical signs experienced by the animals during the experiments and additional factors can influence the prioritization between reduction and refinement.

In **Paper II-III** and **V**, we investigated designs using Monte Carlo simulations. An alternative approach would be to perform a classical optimal design study using a design criterion based on the expected variance of the parameters. Such an investigation, with ethical costs as in **Paper V,** could help limit the otherwise impractically large number of combinations of designs, ethical weights, dose placements and curves that needs to be tested.

In **Paper V** we investigated the impact of ethical weights on the performance of different with quantal data and a similar study with continuous data is warranted.

**Paper V** was a study on the ethical cost-efficiency of a single dose-response study. Nordberg and colleagues (2008) have investigated the monetary cost-efficiency of different tests in relation to the criteria for labelling and classification. Animal welfare could be included in such strategies as well. To do so the ethical cost of different tests (acute, subacute, irritation etc) needs to be estimated. Gabbert and van Ierland (2010) made a similar investigation on

mutagenicity tests comparing the efficiency of *in vitro* and *in vivo* tests. However, their investigation only included number of animals as a proxy for animal welfare. The ethical cost of the different type of in vivo studies, depending on the expected distress of the animals, could be included as a factor as well.

Monetary cost could also be included in the analysis such as the ones in **Paper V,** by setting monetary cost boundaries, for example by setting a limit on the numbers used as well as a limit on the ethical cost of the study.

# 7  ACKNOWLEDGEMENTS

My main supervisor **Mattias Öberg** for accepting me as a PhD student. Your patience, encouragement and ability to always see things positively exceeds anything I have ever encountered before.

My co-supervisor and head of the unit **Gunnar Johanson**, for being the best boss I have ever had. The working climate at our unit is great in large part because you lead it. I would also like to thank you for all your scientific advice.

My mentor **Anders Grahnén,** I have always known that you would be there if I needed you, just as you were for us at Pharmen during my undergraduate studies.

My colleagues and co-authors:  **Salomon Sand**, for really introducing me to the world of the Benchmark Dose. **Fereshteh Kalantari** for all our fruitful discussions about animal distributions and quality metrics. I also owe you thanks for encouraging me to do model averaging. **Elin Törnqvist**, thanks for all your positive energy. Your experience with animals has been absolutely vital for this project and you are also such an excellent team member. **Christina Rudén** and **Sven-Ove Hansson,** for your essential inputs on Paper IV. **Helen Håkansson** and **Maria Herlin**, for collaborations that are not part of this thesis but nonetheless helped me become the scientist I am today.

**Antero da Silva**, my master student in toxicology. Your master thesis work was as important for my development as any of the papers in this thesis. Thank you for putting such effort into a sometimes tedious project.

**Ian Jarvis** and **Joe DePierre** for very valuable language editing of some articles and parts of this thesis.

All the various friends and colleagues who were "guinea pigs" in the interview study and all the participants in the actual study.

Tack alla kollegor och vänner som på olika sätt har förgyllt de senaste 5 åren:

**Mia Johansson**, för att du varit en fantastisk "science sister" och för att du, ofta ganska bokstavligt, lyst upp min doktorandtillvaro. **Ulrika Carlander,** speciellt för all input till modelleringsseminarierna. Tack till er båda för läsningen av tidiga utkast till denna

avhandling. **Aishwarya Mishra**, som har varit med och gått parallellt med mig nästan hela resan från registrering till disputation och såklart alla andra doktorandkollegor som alla bidragit till allt kul: **Afshin Mohammadi Bardboori**, **Anna-Karin Mörk**, **Johanna Bengtsson**, **Kristin Stamyr** och **Stephanie Juran**. Jag saknar er/kommer att sakna er.

**Bengt Sjögren**, en av de vänligaste personerna jag mött. Tack för att du tog mig med i Mundialistas.

Alla orginalarbtoxare (eller arbtoxorginal?): **Agneta Rannug**, **Anne Vonk**, **Annika Calgheborn**, **Anteneh Desalegn**, **Birgit Postol**, **Carolina Vogs**, **Cecilia Wallin**, **Dingsheng Li**, **Emma Wincent**, **Johan Ljungberg**, **Johnny Lorentzen**, **Koustav Ganguly**, **Kristin Larsson**, **Lena Ernstgård**, **Lina Graner**, **Linda Bergander**, **Linda Schenk**, **Marc-Andre Verner**, **Maria Jönsson**, **Martin Fransson**, **Matias Rauma**, **Michail Panagiotakis**, **Ophelie Brenner**, **Ramesh Thapaliya**, **Rosella Dallo**, **Siraz Shaik**, **Tao Liu**, **Tshepo Moto** och **Uriell Deng**. Det har varit kul att arbeta med er under de här åren.

Alla nya **dermatologiska enhetsmedlemmar** samt de nya **korridorsvännerna inom lung- och allergiforskning** och alla på **Swetox** för luncher och fikan. Pseudo-enhetsmedlemmarna på Arbetsmiljöverket: **Anders Iregren**, **Anna-Karin Alexandrie**, **Birgitta Lindell**, **Jill Järnberg** och **Johan Montelius** samt **Marie Nyman** och **Jenny Carlsson** på Gentekniknämnden, för alla diskussioner om gränsvärden, veckans brott, backtrav, storcitrus och lillcitrus.

**Ricardo**, **Gunnar** och alla andra i **Mundialistas** och **Magna Carta**, för fantastisk fotboll. **Stockbowl**, för fantasy football. **LIPS** för helgerna. **Rechoir**, för valborgsfrukostar, medeltidsveckor och för att ni tagit hand om Karin på torsdagarna. **Maria B** för att du är en så bra vän och för allt stöd du och **Marcus B** gett mig. Det har betytt mer än ni tror. **Alla gamla farmisar**, speciellt **Angelica** som tipsade om doktorandplatsen på KI och **Erik A** som illustrerat framsidan på denna avhandling. **Börje-Fredrik**, **Linnea** och **Peter** för såväl verklig som imaginär vänskap. **Martin Styhre**, för att du varit med sen waaaay back och för att du tagit halva förnuftet till fånga och åtminstone flyttat till inom tågreseavstånd.

Min nya extrafamilj: **Hans**, **Gunilla**, **Anne**, **Gunnar** och **Åsa** för alla familjehelger, för att ni bidragit till att Karin blivit som hon är och för att all tänkbar hjälp med allt möjligt rörande hus och barn.

Min gamla vanliga familj: Mina föräldrar **Göran** and **Kjerstin**, som alltid funnits där för mig. Mina syskon **Anneli**, **Jesper** och **Lisa** och deras familjer som varit en stabil klippa av bohusgranit i Stenungsund under alla mina år i självvald exil. **Karin Å** för att du alltid skämmer bort oss och **Karin R** och **Krille** för alla jular.

Allra sist, men inte minst: **Karin L**, tack för dina uppoffringar under de sista veckorna och för att du flyttat med mig överallt dit mina studier fört mig. Du är mitt livs kärlek och jag kan inte föreställa mig livet utan dig. **Elias**, du är den mest fantastiska lilla son man kan tänka sig.

# 8 REFERENCES

Allen, B. C., et al., 1994a. Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels. Fundam Appl Toxicol. 23, 487-95.

Allen, B. C., et al., 1994b. Dose-Response Assessment for Developmental Toxicity. III. Statistical-Models. Fundamental and Applied Toxicology. 23, 496-509.

Allen, B. C., et al., 1996. Benchmark dose analysis of developmental toxicity in rats exposed to boric acid. Fundam Appl Toxicol. 32, 194-204.

Bailer, A. J., et al., 2005a. Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Analysis. 25, 291-299.

Bailer, A. J., et al., 2005b. Incorporating uncertainty and variability in the assessment of occupational hazards. International Journal of Risk Assessment and Management. 5, 344-357.

Balcombe, J. P., 2006. Laboratory environments and rodents' behavioural needs: a review. Laboratory Animals. 40, 217-235.

Baron, J., et al., 2001. Analog scale, magnitude estimation, and person trade-off as measures of health utility: Biases and their correction. Journal of Behavioral Decision Making. 14, 17-34.

Barton, H. A., et al., 1998. Dose-response characteristics of uterine responses in rats exposed to estrogen agonists. Regulatory Toxicology and Pharmacology. 28, 133-149.

Berry, C. L., 1988. The No-Effect Level and Optimal Use of Toxicity Data. Regulatory Toxicology and Pharmacology. 8, 385-388.

Bhattacharya, R., Lin, L. Z., 2010. An adaptive nonparametric method in benchmark analysis for bioassay and environmental studies. Statistics & Probability Letters. 80, 1947-1953.

Blackorby, C., et al., 1997. Critical-Level Utilitarianism and the Population Ethics-Dilemma. Economics and Philosophy. 13, 197-230.

Bogdanffy, M. S., et al., 2001. Harmonization of cancer and noncancer risk assessment: proceedings of a consensus-building workshop. Toxicological Sciences. 61, 18-31.

Brandon, E. F., et al., 2013. Does EU legislation allow the use of the Benchmark dose (BMD) approach for risk assessment? Regul Toxicol Pharmacol. 67, 182-8.

Buckland, S. T., et al., 1997. Model selection: An integral part of inference. Biometrics. 53, 603-618.

Budtz-Jorgensen, E., et al., 2001. Benchmark dose calculation from epidemiological data. Biometrics. 57, 698-706.

Canadian Council on Animal Care, Categories of Invasiveness in Animal Experiments. Ottawa, 1991.

Canadian Council on Animal Care, CCAC Animal Data Report 2013. Ottawa, 2015.

Carlson, E., 1998. Mere Addition and Two Trilemmas of Population Ethics. Ethics and Philosophy. 14, 283-306.

Cohen, J., et al., 2006. European public acceptance of euthanasia: Socio-demographic and cultural factors associated with the acceptance of euthanasia in 33 European countries. Social Science & Medicine. 63**,** 743-756.

Crump, K., 2002. Critical issues in benchmark calculations from continuous data. Critical Reviews in Toxicology. 32**,** 133-153.

Crump, K. S., 1984. A new method for determining allowable daily intakes. Fundam Appl Toxicol. 4**,** 854-71.

Crump, K. S., 1995. Calculation of Benchmark Doses from Continuous Data. Risk Analysis. 15**,** 79-89.

Damschroder, L. J., et al., 2004. The validity of person tradeoff measurements: Randomized trial of computer elicitation versus face-to-face interview. Medical Decision Making. 24**,** 170-180.

Damschroder, L. J., et al., 2007. Why people refuse to make tradeoffs in person tradeoff elicitations: A matter of perspective? Medical Decision Making. 27**,** 266-280.

Dankovic, D., et al., 2007. An approach to risk assessment for TiO2. Inhalation Toxicology. 19**,** 205-212.

Davis, J. A., et al., 2011. Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1. Toxicol Appl Pharmacol. 254**,** 181-91.

Dawkins, M. S., 2008. The science of animal suffering. Ethology. 114**,** 937-945.

de Boo, M. J., et al., 2005. The interplay between replacement, reduction and refinement: considerations where the Three Rs interact. Animal Welfare. 14**,** 327-332.

Dekkers, S., et al., 2001. Critical effect sizes in toxicological risk assessment: a comprehensive and critical evaluation. Environmental Toxicology and Pharmacology. 10**,** 33-52.

Dekkers, S., et al., 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 Anal. 26**,** 867-80.

Dette, H., et al., 2009. Optimal designs for dose-finding experiments in toxicity studies. Bernoulli. 15**,** 124-145.

Dolan, P., Tsuchiya, A., 2003. The person trade-off method and the transitivity principle: an example from preferences over age weighting. Health Economics. 12**,** 505-510.

Dybing, E., et al., 2002. Hazard characterisation of chemicals in food and diet. dose response, mechanisms and extrapolation issues. Food Chem Toxicol. 40**,** 237-82.

EC, Annex VII: Standard information requirements for substances manufactured or imported in quantities of one tonne or more Vol. 2014, 2006a.

EC, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. 2006b.

ECHA, "Guidance on information requirements and chemical safety assessment, Chapter R.8: Characterisation of dose [concentration]–response for human health" In: European Chemicals Agency, (Ed.), Helsinki, Finland., 2012.

ECHA, The Use of Alternatives to Testing on Animals for the REACH Regulation. Second report under Article 117(3) of the REACH Regulation. Helsinki, 2014.

Edler, L., 2014. Benchmark Dose in Regulatory Toxicology. in: Reichl, F.-X., Schwenk, M., (Eds.), Regulatory Toxicology. Springer Verlag, Berlin.

Edler, L., et al., 2002. Mathematical modelling and quantitative methods. Food Chem Toxicol. 40**,** 283-326.

EEC, 1986. Council Directive 86/609/EEC of 24 November 1986 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 of the European Communities. L358**,** 1-29.

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 (2009) 1-72.

EU, 2010. Directive 2010/63/EU of the European Parliament and of the councilL of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European Union. 33-79.

Executive Committee of the Congress, 2000. The Three Rs Declaration of Bologna: Reduction, Refinement and Replacement Alternatives and Laboratory Animal Procedures Adopted by the 3rd World Congress on Alternatives and Animal Use in the Life Sciences, Bologna, Italy, on 31 August 1999. Atla-Alternatives to Laboratory Animals. 28**,** 1-5.

Faes, C., et al., 2007. Model averaging using fractional polynomials to estimate a safe level of exposure. Risk Analysis. 27**,** 111-123.

Franco, N. H., Olsson, I. A. S., 2014. Scientists and the 3Rs: attitudes to animal use in biomedical research and the effect of mandatory training in laboratory animal science. Laboratory Animals. 48**,** 50-60.

Fung, K. Y., et al., 1998. A comparison of methods for estimating the benchmark dose based on overdispersed data from developmental toxicity studies. Risk Anal. 18**,** 329-42.

Gabbert, S., van Ierland, E. C., 2010. Cost-Effectiveness Analysis of Chemical Testing for Decision-Support: How to Include Animal Welfare? Human and Ecological Risk Assessment. 16**,** 603-620.

Gaylor, D., et al., 1998. Procedures for calculating benchmark doses for health risk assessment. Regulatory Toxicology and Pharmacology. 28**,** 150-164.

Gaylor, D. W., Slikker, W., 1990. Risk Assessment for Neurotoxic Effects. Neurotoxicology. 11**,** 211-218.

Guha, N., et al., 2013. Nonparametric Bayesian Methods for Benchmark Dose Estimation. Risk Analysis. 33**,** 1608-1619.

Hajar, R., 2011. Animal testing and medicine. Heart Views. 12**,** 42.

Hansen, A. K., et al., The need to refine the notion of reduction., Humane Endpoint in Animal Experiments for Biomedical Research. Laboratory Animals Ltd, London, 1999.

Hansen, L. A., 2013. Institution animal care and use committees need greater ethical diversity. Journal of Medical Ethics. 39, 188-190.

Holland-Letz, T., Kopp-Schneider, A., 2015. Optimal experimental designs for dose-response studies with continuous endpoints. Archives of Toxicology. 89, 2059-2068.

Kavlock, R. J., et al., 1995. Dose-response assessments for developmental toxicity. IV. Benchmark doses for fetal weight changes. Fundam Appl Toxicol. 26, 211-22.

Kavlock, R. J., et al., 1996. A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. Risk Analysis. 16, 399-410.

Keating, S. C. J., et al., 2012. Evaluation of EMLA Cream for Preventing Pain during Tattooing of Rabbits: Changes in Physiological, Behavioural and Facial Expression Responses. PLoS One. 7.

Knight, A., 2013. The Costs and Benefits of Animal Experiments. Palgrave Macmillan, Basingstoke, Hampshire, UK.

Kodell, R. L., et al., 1991. Mathematical-Modeling of Reproductive and Developmental Toxic Effects for Quantitative Risk Assessment. Risk Analysis. 11, 583-590.

Kolbe, T., et al., 2015. Lifetime Dependent Variation of Stress Hormone Metabolites in Feces of Two Laboratory Mouse Strains. PLoS One. 10, e0136112.

Krewski, D., et al., 2002. Optimal designs for estimating the effective dose in developmental toxicity experiments. Risk Analysis. 22, 1195-1205.

Kuljus, K., et al., 2006. Comparing experimental designs for benchmark dose calculations for continuous endpoints. Risk Anal. 26, 1031-43.

Langford, D. J., et al., 2010. Coding of facial expressions of pain in the laboratory mouse. Nat Methods. 7, 447-9.

Matsuura, T., et al., 2012. Estimation of restraint stress in rats using salivary amylase activity. J Physiol Sci. 62, 421-7.

Meek, M. E., et al., 2002. Guidelines for application of chemical-specific adjustment factors in dose/concentration - response assessment. Toxicology. 181, 115-120.

Moerbeek, M., et al., 2004. A comparison of three methods for calculating confidence intervals for the benchmark dose. Risk Anal. 24, 31-40.

Moon, H., et al., 2005. Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness. Risk Analysis. 25, 1147-1159.

Morales, K. H., et al., 2006. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. Journal of the American Statistical Association. 101, 9-17.

Morton, D. B., Griffiths, P. H. M., 1985. Guidelines on the Recognition of Pain, Distress and Discomfort in Experimental-Animals and an Hypothesis for Assessment. Veterinary Record. 116, 431-436.

Murray, C. J. L., Lopez, A. D., The global burden of disease. World Health organization, Harvard School of Public Health, World Bank, Geneva, 1996.

Murrell, J. A., et al., 1998. Characterizing dose-response: I: Critical assessment of the benchmark dose concept. Risk Anal. 18, 13-26.

NAC/AEGL, Standing Operating Procedures for Developing Acute Exposure Guideline Levels for Hazardous Chemicals. In: Subcommittee on Acute Exposure Guideline Levels, N. R. C., (Ed.), Washington, DC, 2001.

Ng, Y.-K., 1989. What should we do about future generations. Economics and Philosophy. 5, 235-253.

Nord, E., 1995. The Person-Trade-Off Approach to Valuing Health-Care Programs. Medical Decision Making. 15, 201-208.

Nordberg, A., et al., 2008. Towards more efficient testing strategies - Analyzing the efficiency of toxicity data requirements in relation to the criteria for classification and labelling. Regulatory Toxicology and Pharmacology. 50, 412-419.

NRC, Risk Assessment in the Federal Government: Managing the Process. 1983.

NTP, Toxicology and Carcinogenesis - Studies of Furan (CAS No. 110-00-9) in F344/n Rats and B6C3F1 Mice (Gavage studies) In: U.S. Department of Health and Human Services, (Ed.), 1993.

Öberg, M., 2010. Benchmark dose approaches in chemical health risk assessment in relation to number and distress of laboratory animals. Regul Toxicol Pharmacol. 58, 451-4.

OECD, Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Human Endpoints for Experimental Animals Used in Safety Evaluation. 2002.

OECD, OECD Guidelines for the Testing of Chemicals 2012.

Olsson, A. S., et al., 2012. The 3Rs principle - mind the ethical gap. ALTEX Proceedings. 1/12, 333-336.

Piegorsch, W. W., et al., 2013. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. Environmetrics. 24, 143-157.

Piegorsch, W. W., et al., 2012. Nonparametric estimation of benchmark doses in environmental risk assessment. Environmetrics. 23, 717-728.

Porter, D. G., 1992. Ethical Scores for Animal-Experiments. Nature. 356, 101-102.

Rai, K., Vanryzin, J., 1985. A Dose-Response Model for Teratological Experiments Involving Quantal Responses. Biometrics. 41, 1-9.

Regan, T., 1983. The Case for Animal Rights. Routledge & Kegan Paul, London.

Ringblom, J., et al., 2014. Current modeling practice may lead to falsely high benchmark dose estimates. Regulatory Toxicology and Pharmacology. 69, 171-177.

Rollin, B. E., 2006. The regulation of animal research and the emergence of animal ethics: A conceptual history. Theoretical Medicine and Bioethics. 27, 285-304.

Rovida, C., Hartung, T., 2009. Re-Evaluation of Animal Numbers and Costs for In Vivo Tests to Accomplish REACH Legislation Requirements for Chemicals - a Report by the Transatlantic Think Tank for Toxicology (t(4)). Altex-Alternativen Zu Tierexperimenten. 26, 187-208.

Russell, W. M. S., Burch, R. L., 1959. The principles of humane experimental technique. Methuen, London.

Sand, S., et al., 2011. A Signal-to-Noise Crossover Dose as the Point of Departure for Health Risk Assessment. Environmental Health Perspectives. 119, 1766-1774.

Sand, S., et al., 2008. The current state of knowledge on the use of the benchmark dose concept in risk assessment. J Appl Toxicol. 28, 405-21.

Sand, S., et al., 2006. Identification of a critical dose level for risk assessment: developments in benchmark dose analysis of continuous endpoints. Toxicological Sciences. 90, 241-51.

Sandøe, P., Christiansen, S. B., 2007. The value of animal life: how should we balance quality against quantity? Animal Welfare. 16, 109-115.

Scharmann, W., 1999. Physiological and ethological aspects of assessment of pain, distress and suffering. in: Hendriksen, C. F. M., Morton, D. B., (Eds.), In Humane endpoints in animal experiments for biomedical research. Royal Society of Medicine Press, London, pp. 33-39.

Schwarzinger, M., et al., 2004. Lack of multiplicative transitivity in person trade-off responses. Health Economics. 13, 171-181.

Shao, K., Gift, J. S., 2014. Model Uncertainty and Bayesian Model Averaged Benchmark Dose Estimation for Continuous Data. Risk Anal. 34, 101-120.

Shao, K., et al., 2013. Is the assumption of normality or log-normality for continuous response data critical for benchmark dose estimation? Toxicology and Applied Pharmacology. 272, 767-779.

Shao, K., Small, M. J., 2012. Statistical Evaluation of Toxicological Experimental Design for Bayesian Model Averaged Benchmark Dose Estimation with Dichotomous Data. Human and Ecological Risk Assessment. 18, 1096-1119.

Simmons, S. J., et al., 2015. Bayesian model averaging for benchmark dose estimation. Environmental and Ecological Statistics. 22, 5-16.

Singer, P., 2009. Animal Liberation. HarperCollins Publishers, New York.

Slob, W., 1999. Thresholds in toxicology and risk assessment. International Journal of Toxicology. 18, 259-268.

Slob, W., 2002. Dose-response modeling of continuous endpoints. Toxicological Sciences. 66, 298-312.

Slob, W., 2007. What is practical threshold? Toxicologic Pathology. 35, 848-849.

Slob, W., 2014a. Benchmark dose and the three Rs. Part I. Getting more information from the same number of animals. Crit Rev Toxicol. 1-11.

Slob, W., 2014b. Benchmark dose and the three Rs. Part II. Consequences for study design and animal use. Crit Rev Toxicol. 1-13.

Slob, W., et al., 2005. A statistical evaluation of toxicity study designs for the estimation of the benchmark dose in continuous endpoints. Toxicological Sciences. 84, 167-85.

Slob, W., Pieters, M. N., 1998. A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: General framework. Risk Analysis. 18, 787-798.

Slob, W., Setzer, R. W., 2014. Shape and steepness of toxicological dose-response relationships of continuous endpoints. Crit Rev Toxicol. 44, 270-297.

Solecki, R., et al., 2005. Guidance on setting of acute reference dose (ARfD) for pesticides. Food Chem Toxicol. 43, 1569-93.

Sotocinal, S. G., et al., 2011. The Rat Grimace Scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. Molecular Pain. 7.

Spielmann, H., et al., 2011. A Critical Evaluation of the 2011 ECHA Reports on Compliance with the REACH and CLP Regulations and on the Use of Alternatives to Testing on Animals for Compliance with the REACH Regulation. Atla-Alternatives to Laboratory Animals. 39, 481-493.

Stafleu, F. R., et al., 1999. The ethical acceptability of animal experiments: a proposal for a system to support decision-making. Laboratory Animals. 33, 295-303.

Tännsjö, T., 2002. Why We Ought to Accept the Repugnant Conclusion. Utilitas. 14, 339-359.

Taylor, K., et al., 2008. Estimates for worldwide laboratory animal use in 2005. Atla-Alternatives to Laboratory Animals. 36, 327-342.

Torrance, G. W., 1986. Measurement of Health State Utilities for Economic Appraisal - a Review. Journal of Health Economics. 5, 1-30.

Travis, K. Z., et al., 2005. The role of the benchmark dose in a regulatory context. Regul Toxicol Pharmacol. 43, 280-91.

Ubel, P. A., et al., 2001. Preference for equity as a framing effect. Medical Decision Making. 21, 180-189.

Ubel, P. A., et al., 1996. Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. Medical Decision Making. 16, 108-116.

Ubel, P. A., et al., 2002. Exploring the role of order effects in person trade-off elicitations. Health Policy. 61, 189-199.

US Department of Agriculture, Animal Welfare Act. 2013.

US EPA, A Review of the Reference Dose and Reference Concentration Processes. Washington, 2002.

US EPA, Benchmark dose technical guidance document. EPA/100/R-12/001. . In: U.S. Environmental Protection Agency, R. A. F., (Ed.), Washington, DC: , 2012.

US EPA, Background and Support Materials for Peer Consultation Webinar Workshop on Model Averaging Methods for Dose-Response Analysis. In: Development, O. o. R. a., (Ed.), Washington, 2015.

US. EPA, Guidelines for Carcinogen Risk Assessment In: Risk Assessment Forum, U. S. E. P. A., (Ed.), Washington, 2005.

USEPA, The use of the benchmark dose (BMD) approach in health risk assessment. Final report. EPA/630/R-94/007. . In: Risk Assessment Forum, U. S. E. P. A., (Ed.), Washington, DC., 1995.

Venzon, D. J., Moolgavkar, S. H., 1988. A Method for Computing Profile-Likelihood-Based Confidence-Intervals. Applied Statistics-Journal of the Royal Statistical Society Series C. 37, 87-94.

Vieira de Castro, A. C., Olsson, A. S., 2015. Does the Goal Justify the Methods? Harm and Benefit in Neuroscience Research Using Animals. in: Lee, G., et al., (Eds.), Ethical Issues in Behavioral Neuroscience. vol. 19. Springer, Berlin.

Wang, K., et al., 2013. Two-Stage Experimental Design for Dose-Response Modeling in Toxicology Studies. Acs Sustainable Chemistry & Engineering. 1, 1119-1128.

Weller, E. A., et al., 1995. Implications of Developmental Toxicity Study Design for Quantitative Risk Assessment. Risk Analysis. 15, 567-574.

West, R. W., Kodell, R. L., 1999. A comparison of methods of benchmark-dose estimation for continuous response data. Risk Anal. 19, 453-9.

West, R. W., et al., 2012. The impact of model uncertainty on benchmark dose estimation. Environmetrics. 23, 706-716.

Wheeler, M., Bailer, A. J., 2012. Monotonic Bayesian Semiparametric Benchmark Dose Analysis. Risk Anal. 32, 1207-1218.

Wheeler, M. W., Bailer, A. J., 2007. Properties of model-averaged BMDLs: A study of model averaging in dichotomous response risk estimation. Risk Analysis. 27, 659-670.

Wheeler, M. W., Bailer, A. J., 2009a. Benchmark dose estimation incorporating multiple data sources. Risk Anal. 29, 249-56.

Wheeler, M. W., Bailer, A. J., 2009b. Comparing model averaging with other model selection strategies for benchmark dose estimation. Environmental and Ecological Statistics. 16, 37-51.

WHO, Principles for the Assessment of Risks to Human Health from Exposure to Chemicals. Envrional Health Criteria 210, Geneva, 1999.

WHO, Principles for modeling dose–response for the risk assessment of chemicals. In: Environmental Health Criteria 239, (Ed.), Geneva, 2009.

WHO/IPCS, Harmonization document 1. IPCS risk assessment terminology. . In: Organization, W. H., (Ed.), Geneva, 2004.