Research Article

Assigning Ethical Weights to Clinical Signs Observed During Toxicity Testing

Joakim Ringblom[1], Elin Törnqvist[1,2], Sven Ove Hansson[3], Christina Rudén[4] and Mattias Öberg[1,2]

[1]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; [2]Swedish Toxicology Sciences Research Center (Swetox), Södertälje, Sweden; [3]Royal Institute of Technology, Department of Philosophy and History, Stockholm, Sweden; [4]Stockholm University, Department of Environmental Science and Analytical Chemistry, Stockholm, Sweden

Summary
Reducing the number of laboratory animals used and refining experimental procedures to enhance animal welfare are fundamental questions to be considered in connection with animal experimentation. Here, we explored the use of cardinal ethical weights for clinical signs and symptoms in rodents by conducting trade-off interviews with members of Swedish Animal Ethics Committees in order to derive such weights for nine typical clinical signs of toxicity. The participants interviewed represent researchers, politically nominated political nominees and representatives of animal welfare organizations. We observed no statistically significant differences between these groups with respect to the magnitude of the ethical weights assigned, though the political nominees tended to assign lower weights. Overall, hunched posture was considered the most severe clinical sign and body weight loss the least severe. The ethical weights assigned varied considerably between individuals, from zero to infinite value, indicating discrepancies in prioritization of reduction and refinement. Cardinal ethical weights may be utilized to include both animal welfare refinement and reduction of animal use in designing as well as in retrospective assessment of animal experiments. Such weights may also be used to estimate ethical costs of animal experiments.

Keywords: 3Rs, animal ethics, ethical weights, ethical committees, toxicity testing

1 Introduction

In 1959 Russell and Burch published their milestone book entitled "The Principles of Humane Experimental Technique", in which they introduced the 3Rs (replacement, reduction, refinement) that continue to have a major influence on regulations and ethical principles regarding research involving animals. The goals of the 3Rs are to replace animal tests with animal-free methods whenever scientifically justifiable, to reduce the number of animals while maintaining a reliable scientific outcome, and to use refined methods that minimize or eliminate potential pain, suffering or distress experienced by experimental animals (Russell and Burch, 1959).

Of the 11.5 million animals utilized for experimental and other scientific purposes in the EU during 2011, approximately 1 million were used in toxicological studies or safety evaluations (EC, 2013). Such investigations are regulated by laws and guidelines, and if there are no alternative methods, toxicity testing on animals is required to enable valid risk assessment of chemicals. All research based on animal methods in the EU is regulated by Directive 2010/63/EU on the protection of animals used for scientific purposes, which has been transposed into national law in the member states and in which the importance of the 3R principles is emphasized. The 3R principles are also advocated by various legislations related to chemical safety. The EU's regulation on chemicals, REACH (EC, 2006), for example, promotes the use of alternative methods for animal testing, but does not oblige the test performer to do so (Article 25.1). It is no longer permitted to use animal based methods for toxicity testing of cosmetic products or ingredients in Europe (EC, 2009).

Several studies have been made on strategies to obtain as much toxicological information as possible, given a fixed number of animals (Shao and Small, 2012; Slob, 2014) or a fixed monetary cost (Nordberg et al., 2008). However, none of these studies takes animal suffering into account although ethical cost is also an important consideration in such an assessment (Öberg, 2010).

The ethical costs of animal experimentation have several components. Unquestionably, one central component is suffering or, more generally, negative experiences of individual animals. Appraisal of the suffering incurred in any given experiment will have to include both the suffering of each individual animal and the number of animals subjected to this level of suffering. However, it is not obvious how the number of animals and the degree of distress should be weighed against each other, and there is a lack of guidance on prioritizing between the 3Rs when they are in conflict, e.g., how to trade-off a reduction of animal numbers against higher levels of suffering per animal (de Boo et al., 2005). A report from the Nordic-European Workshop on Ethical Evaluation of Animal Experiments maintained that more emphasis should be placed on the level of suffering than on the number of animals (Voipio et al., 2004), and similarly the Swiss Academy of Medical Sciences (SAMS) stated:

"If the suffering of individual animals can be reduced significantly through the use of a larger number of animals, the reduction of individual suffering shall take priority over the reduction of the number of animals used in the experiment" (SAMS, 2005).

However, it remains unclear by how much suffering must be reduced for this to apply. Further, the evaluation of the extent of suffering based on various signs and degrees of distress requires expert judgment and cannot always be estimated correctly during design of a study.

Laboratory animals are bred for a specific purpose. Thus, laboratory animals "saved" by choosing an experimental set-up with fewer animals will not be released, but simply used in another experiment. In the long run, choosing a setup with fewer animals will lead to fewer experimental animals being brought into existence. Sandøe and Christiansen explored the quality and quantity of life concept by discussing differences in how quality is weighed against quantity for animals bred for science and food with pets and humans (Sandøe and Christiansen, 2007). Decisions about euthanizing an animal by a veterinarian or a researcher or, more controversially, a terminally ill human patient by a doctor, may reflect a view where only quality of life matters, while other views also include quantity of life, i.e., consider that life itself has a value. Related issues regarding the quantity and quality of life of both farm animals and laboratory animals and the consequences of different ethical perspectives have also been discussed in more detail by Franco and colleagues (2014).

The question of individual distress versus the number of animals can also be viewed as analogous to the person trade-off questions designed to determine disability-adjusted life years (DALY) or quality adjusted life years (QALY) for humans (Gold et al., 2002). In these assessments, trade-off respondents are asked, e.g., if they would choose to cure a few patients whose health is very poor or a larger number in somewhat better health, given that the cost of treatment for both options is the same (Nord, 1995).

The use of animals in research, safety assessment, and education is regulated by national laws worldwide. Weighing suffering against benefit, a predominantly utilitarian perspective, is currently the most common approach to determine the acceptability of animal use (Vieira de Castro and Olsson, 2015). Utilitarianism is the ethical theory that assesses alternatives for human action on the basis of the total positive effects minus the total negative effects of each alternative. As an ethical theory, it is by no means uncontroversial. Its application to human welfare has been criticized as being insensitive to the distribution of welfare as well as to issues such as rights, consent and intentionality. On the other hand, utilitarianism is remarkably adaptable and it has also been subject to the opposite critique of being so capable of different interpretations that it is almost empty of content (Hansson, 2014). However, in general terms, utilitarianism is an obvious reference point in moral philosophy. For our present purposes, we will therefore assume that considerations of animal welfare have a utilitarian structure.

Several laws and guidelines provide guidance for the assessment and classification of animal harm. There are similarities in the types of classification, mostly mild to severe, and in how types of experimental procedures causing pain, suffering or distress are classified in different countries (Purves, 2000). In the EU Directive on the protection of animals used for scientific purposes (EU, 2010) examples of mild, moderate and severe experimental procedures are listed, and at implementation of this legislation in 2013, a detailed guidance document (EC, 2010) illustrating how to set up processes of severity classification, day-to-day assessment and actual severity assessment in different research areas was developed. It aims at refining the experimental situation for research animals, assessed by clinical observations, e.g., body weight, coat condition, body function, behavior and locomotion, and appropriate endpoints. Similar clinical scoring sheets were developed previously for the assessment of pain, distress and suffering of individual animals (Morton and Griffiths, 1985; Scharmann, 1999) and groups of animals (Leach et al., 2008).

However, to estimate the total ethical cost of experimental animals and weigh it quantitatively against the scientific gain, cardinal measures with ethical weights on a ratio scale are required. Such a measure should have the following two key properties: Firstly, the value assigned to the ethical cost of using an individual animal should be proportional to the ethical severity of that use. For instance, if a measurement of animal suffering yields an ethical weight of 6 in one case and the value 3 in another case, then the suffering should be judged to be twice as severe in the one case in comparison to the other. Secondly, a practicable measure should be additive with respect to individual animals. In other words, the severity of exposing several animals to distress should be equal to the sum of the severities for each individual animal, i.e., when a group of animals is exposed to the same treatment, the severity is proportional to the number of animals in the group.

The main objective of the current investigation was to explore the assignment of cardinal ethical weights to clinical signs in animal experiments and to contribute to the development of more efficient harm-benefit analyses in this context. In addition, we investigated how groups of laypersons and researchers perceive and rate different clinical signs of toxicity in experimental animals.

2 Methods

2.1 Description of Swedish Animal Ethics Committees and selection of test participants

To receive ethical approval for an experiment involving animals, the potential benefit of the experiment must outweigh the harm to the animals, according to EU Directive 2010/63/EU (EU, 2010). In Sweden, applications for such an approval are assessed by regional Animal Ethics Committees, each led by a judge and composed of twelve other members including researchers (6), politically nominated laypersons (4), and representatives of animal welfare organizations (2). Each member has a personal alternate member. In total, there are 230 members and alternate members. 217 of these were invited by e-mail to take part in a telephone interview. Twelve members were not contacted, either because we were unable to retrieve their contact information or because the position was vacant at the time. One member was excluded because she had previously provided advice on the design of this study.

The participants were informed that the overriding goal of our project was to improve the situation of animals in toxicological tests while, at the same time, maintaining or even improving the scientific quality and that an important aspect was to determine how symptoms in experimental animals are valued ethically and graded by different individuals. The participants received no compensation for their participation and were informed that they could withdraw at any time.

2.2 Interview participants

Fifty-five committee members agreed to be interviewed. Eight of them (5 researchers, 1 politically nominated layperson and 2 laypersons nominated by animal welfare organizations), found it too difficult to make the trade-offs and withdrew their participation during the interview. The remaining 47 participants (26 men and 21 women, average age 55 years) consisting of 22 researchers, 18 political nominees and 7 representatives of animal welfare organizations completed the interview (average duration 28 min). None of the judges agreed to be interviewed.

2.3 Trade-off scenarios

The participants' views on the relative ethical costs of animal experiments were evaluated on the basis of trade-off questions, similar to person trade-off (PTO) questions employed in healthcare to determine quality of life (Nord, 1995; Prades, 1997). They were given the following information (originally in Swedish as presented in the supplementary materials, https://doi.org/10.14573/altex.1512211s) about a fictitious animal test:

Imagine a seven-day animal study performed on eight-week-old rats. On one occasion during the study, the rats display a clinical sign. At the end of the study, they are euthanized by inhalation overdose of anesthesia or by sedation with inhalation anesthesia followed by exposure to carbon dioxide. The rats are housed in groups of two or three in polycarbonate cages on a woodchip bedding. The cages are enriched with a plastic nest box, paper nesting material and wooden sticks for gnawing, and the animals have free access to food and water.

The participants were then introduced to the trade-off methodology and asked to compare the ethical cost of various clinical signs exhibited by an animal under the described circumstances (Tab. 1). All the clinical signs and their durations were chosen in consultation with laboratory animal veterinarians and toxicological scientists, on the basis that they may potentially occur in toxicological studies and that they are probably dose dependent. Any two clinical signs designated as "severe" are not necessarily equally severe, but they are more likely to occur at higher doses than their milder counterparts. The participants were asked specifically to only consider animal ethics when answering the questions and to ignore other factors like financial cost or scientific value.

The trade-off questions concerned nine pairs of clinical signs, one mild and one severe, the latter being more intense and/or lasting longer. The relative ethical cost of having animals experience the mild, severe or no clinical sign at all were evaluated using trade-off questions that were framed as follows:

(a) *How many animals experiencing [mild tremor during 4 hours] would entail the same ethical cost as 10 animals experiencing [severe tremor for 4 hours]?*

(b) *How many animals with no clinical sign would entail the same ethical cost as 10 animals experiencing [mild tremor]?*

(c) *Your answers would imply that*

$$\frac{Answer\ (a)\ \times Answer\ (b)}{10}$$

(calculated by the interviewer) animals with no clinical signs would entail the same ethical cost as 10 animals experiencing [severe tremor for 4 hours]. Is that a satisfactory number or do you wish to change answer (a) or (b)?

During the interview these same three questions were repeated for each of the nine pairs of clinical signs. At the end of the interview, all of the answers to question (c) were repeated again to give the participants the opportunity to change their answers if they had changed their opinions concerning any of the clinical signs during the interview.

2.4 Data analysis

All answers were converted to ethical weights, with one animal showing no clinical sign being assigned a weight of 1. Thus, if 25 animals with no clinical sign have the same ethical cost as 10 with mild tremor, the ethical weight of mild tremor would be 2.5. Accordingly, the higher the weight the more serious a clinical sign is considered to be and an interviewee who assigned higher weights in general favors refinement over reduction.

A descriptive analysis of the data was performed and the weights assigned by researchers, political nominees and representatives of animal welfare organizations were compared employing Kruskal-Wallis tests with the kruskal.test function in the R package stats (R Core Team, 2013) with a p-value of < 0.05 being considered statistically significant.

Tab. 1: The clinical signs used in the interviews
Each type of clinical sign has one mild and one more severe variant.

	Mild clinical sign	Severe clinical sign
Ataxia	The animal has a loss of ability to coordinate voluntary movements, resulting in walking difficulties and unsteady movements. The animal is still able to eat and drink and perform its natural behavior. Duration 4 h.	The animal has a loss of ability to coordinate voluntary movements, resulting in walking difficulties and unsteady movements. The ataxia affects the ability to perform natural behavior, e.g., reduced food and water consumption could be observed until the clinical sign declines. Duration 4 h.
Tremor	The animal shakes/shivers. The animal is still able to eat and drink and perform its natural behavior. Duration 4 h.	The animal shakes/shivers. The shivering affects the ability to perform natural behavior, e.g., reduced food and water consumption could be observed until the clinical sign declines. Duration 4 h.
Occasional spasm or short convulsion	The animal has occasional involuntary and abnormal muscular contraction in one located muscle group. The animal is conscious during the spasm. Duration: Up to 10 sec.	The animal has convulsions in all parts of the body. The animal is unconscious during the convulsions and i s not affected afterwards. The convulsions are equivalent to a short epileptic seizure. Duration: Up to 10 sec.
Hunched posture	The back is abnormally arched in a concave manner, usually associated with poor condition. Duration: 30 min.	The back is abnormally arched in a concave manner, usually associated with poor condition. Duration: 4 h.
Motor activity decreased	The animal's activity, e.g., movement, is decreased compared to normal. The animal becomes active following stimulation such as noise or physical interaction with animal technician. Duration: 30 min.	The animal's activity, e.g., movement is decreased compared to normal. The animal becomes active following stimulation such as noise or physical interaction with animal technician. Duration: 4 h.
Respiratory clinical signs	The animal breathes slightly faster, slower or more irregularly compared to normal. Duration: A couple of min.	The animal has breathing difficulties with forced breathing. The animal breathes through the mouth instead of the nose. Duration: A couple of min.
Weight loss	5% weight loss during one week.	10% weight loss during one week.
Piloerection	The animal has fur standing on end. Often observed in animals with deteriorated condition and/or when cold. Duration: 30 min.	The animal has fur standing on end. Often observed in animals with deteriorated condition and/or when cold. Duration: 4 h.
Porphyria	The animal has red discharge from the nose and eyes, usually related to stress and/or disease. The discharge (red tear fluid, not blood) results in staining of the fur. Duration: 30 min.	The animal has red discharge from the nose and eyes, usually related to stress and/or disease. The discharge (red tear fluid, not blood) results in staining of the fur. Duration: 4 h.

3 Results

As can be seen in Figure 1, the weights assigned differed widely among the participants. One researcher and one political nominee assigned a weight of 1 to all of the clinical signs. In addition, three researchers, one political nominee and two representatives of animal welfare organizations gave infinite weight to all of these signs.

The Kruskal-Wallis analysis revealed no statistically significant differences between the ratings of the groups with respect to any of the clinical signs. There was however a non-significant trend that the political nominees assigned lower weights than members of the other two groups, with median weights in the range of 2.0-3.5 for mild and 4.0-11 for severe clinical signs. The corresponding values assigned by researchers were somewhat higher, in the range of 2.5-8.0 and 9.5-32, while those for the representatives of animal welfare organizations were 2.0-10 and 8.0-50, respectively.

Of the clinical signs, hunched posture was assigned the highest weight in general with median weights of 4.0 for the mild (duration 30 min) and 20 for the severe clinical sign (duration of 4 h) variant, followed by decreased motor activity with median

weights of 4.0 and 16 for the mild (duration 30 min) respectively severe (duration 4 h) clinical sign (Fig. 1). Weight loss was assigned the lowest median weights, i.e., 2.0 as the median weight for mild (-5% in one week) and 5.3 for severe weight loss (-10% in one week). In general, the mild signs were perceived as entailing similar ethical costs with median weights between 2.0 and 4.0, whereas there were larger differences ranging from median weights between 5.3 and 20 for the signs classified as severe.

4 Discussion

We interviewed members of the Swedish Animal Ethics Committees to see how they evaluate the relative ethical costs of different clinical signs that can occur in laboratory animals during regulatory toxicity testing. This group consisted of researchers, political nominees and representatives of animal welfare organizations with experience in weighing animal suffering against
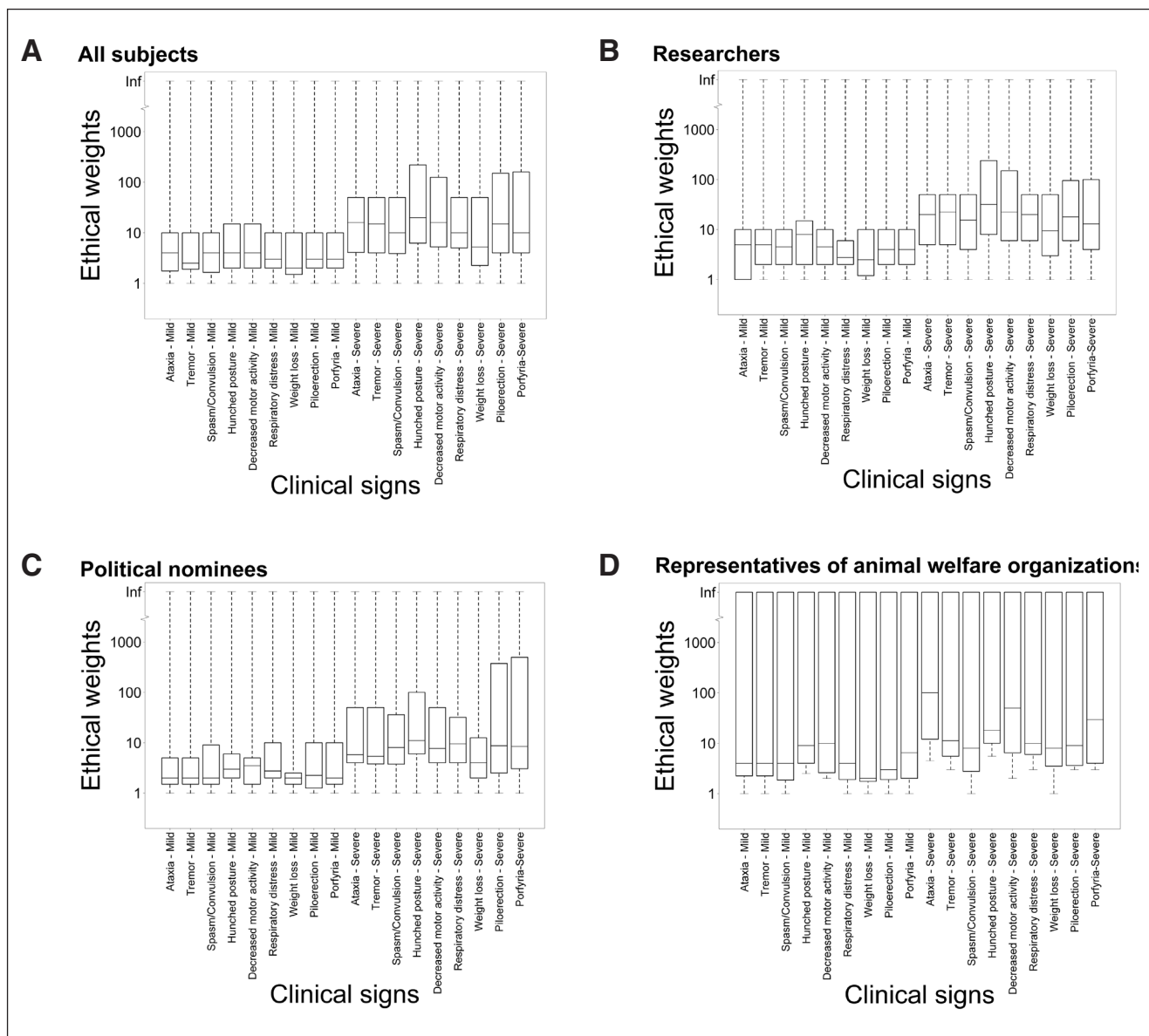
Fig. 1: Boxplot of the distribution of ethical weights assigned to each clinical sign by all 47 participants, the 22 researchers, the 18 political nominees and the 7 representatives of animal welfare organizations
The black line indicates the median weights, the lower and upper end of the boxes the 25th and 75th percentiles, respectively, and the whiskers the minimal and maximal values, respectively. An ethical weight of 10 means that one animal experiencing that sign under the circumstances described entails the same ethical cost as 10 animals free from clinical signs. The infinite maximal values reflect the fact that certain participants felt that it is always better to allow more animals to experience less distress than fewer animals to experimence more distress, regardless of the magnitude of distress. The exact weights can be found in the supplementary materials (https://doi.org/10.14573/altex.1512211s).

scientific gain. The composition of the group of interviewees was fairly similar to that of the committees. It included participants from all regional committees, and both genders were represented. Eight (of the original 55) participants found it too difficult to answer the questions and therefore dropped out of the interview.

In general, representatives of animal welfare organizations and researchers tended to assign higher ethical weights than the political nominees, although these differences were not statistically significant. The ethical weights assigned by the different participants varied greatly, some assigning a weight of 1 and others infinite weight to all clinical signs. There appears to be no strong association between assigning such extreme weights and belonging to a specific group of committee members. Indeed, the ethical views differed more within the groups than between them. While the clinical signs were defined and described in the same manner to all participants, they may still have been interpreted differently, resulting in different ethical weights. However, since members of the ethical committees routinely discuss clinical signs and animal welfare with one another, the differences in their interpretations are likely to be smaller than if one asked the general public.

Two of the 47 participants, one researcher and one political nominee, consistently valued the rats displaying no, mild or severe clinical signs as entailing the same relative ethical cost, i.e., in their opinion no difference in symptoms is large enough to outweigh the increase in ethical cost from using more animals. From this perspective, the number of animals in a study always determines its ethical cost. All the other participants accepted at least to some extent that the use of more animals could reduce the ethical cost if the degree of distress per animal was lower.

Six participants assigned infinite weights to all clinical signs. They also said that it is the experience of the individual animal that counts and therefore it is always better to have more animals experiencing little distress than fewer animals experiencing more distress. From this perspective, the ethical cost of a study would always be determined by the animals that are worst off. This viewpoint corresponds roughly to the Rawlsian viewpoint on the welfare of human societies, according to which the welfare in a country can be judged on the basis of the situation of inhabitants who are worst off (Rawls, 1971).

Thus, in total, 16 out of the 55 interviewees (29%) either refused to trade off or gave ethical weights of 1 or infinity. This is in line with results from applications with the person trade-off approach to decisions concerning the health of humans. In such a study investigating the phenomenon, 22-47% refused to trade off or gave off-scale answers depending on the question and how it was framed (Damschroder et al., 2007).

The balance between reduction and refinement also was recently investigated by Franco and Olsson (2014). In their study, 206 students taking a course in Laboratory Animal Science were asked whether they would prefer to perform a considerably stressful and painful experiment with no permanent effects 20 times on one mouse or once on 20 different mice. A slight majority preferred the second scenario. For animals with a higher sentience or status (chimpanzees or dogs), reduction was considered

more important in relation to refinement than for mice or rabbits. This difference in ethical weights assigned to different species is in need of further empirical investigation.

In the present study, body weight loss was assigned the lowest ethical weight. The limits for acceptable body weight loss were recently challenged in short term studies (up to 7 days) designed to determine the maximum tolerated dose (MTD) for subsequent regulatory testing of the toxicity of pharmaceuticals. The authors stated that a limit at a 10% body weight loss is scientifically and ethically more accurate than the previous practice of accepting body weight losses of over 20% with consequently higher degrees of animal suffering and an exceeded MTD (Chapman et al., 2013). The low rating of weight loss in our study, with the median ethical weight for severe weight loss of 10% only at 5.3, might be anthropomorphic since many humans would not mind losing a few kilos of weight themselves.

The ethical costs of tremor, piloerection, porphyria and convulsion, all representing clinical signs of systemic toxicity and general malaise, were scored with an ethical mean weight between 10 and 16. Hunched posture, which might be more related to pain than to general malaise, was rated highest by the interviewees and scored a mean ethical weight of 20. These severe signs, including body weight loss, are placed in the category "substantial severity" for dose setting in toxicity testing, and doses giving rise to these should be avoided due to risks of too negative effects on animal welfare (NC3R/LASA, 2009). There are no points or numbers related to the clinical signs in the LASA guide, but the signs are grouped into categories with different severity grades (mild, moderate and substantial), and there is quantitative guidance where one sign in the substantial category equals combined signs in the moderate category.

In a similar way, guidance to determine humane endpoints based on severity categories and points is used for decision making regarding pre-terminal euthanasia in animal research in Sweden (Karolinska Institutet, 2016). According to this guidance a mild clinical sign equals 0.1 points and a severe sign equals 0.4 points. When an individual animal reaches a predefined sum of points, usually 0.4 depending on the type of study, the humane endpoint is reached. While these assessment tools are quantitative, or at least semi-quantitative, they are designed for assessing the health of an individual animal and not designed for evaluation of reduction versus refinement.

Various other systems of categorizing the suffering of laboratory animals according to severity have been proposed in the scientific literature. Gradings of the severities of specific experimental procedures as well as assessments of clinical and behavioral changes in relation to duration are commonly weighed against the scientific benefit of the animal use (Stafleu et al., 1999; Delpire et al., 1999; Purves, 2000; Porter, 1992). None of these hold the cardinality property (i.e., proportionality) that is needed to fulfill the needs outlined here. For instance, in the scheme proposed by Stafleu and colleagues, the use of 10 animals costs 0 points, using 10-100 costs 1 point, and using more than 100 animals costs 2 points. Systems as this one are problematic in that a considerable reduction in the number of animals subjected to a treatment (e.g., a reduction from 90 to

12 animals) does not reduce the measure of the total ethical cost.

Several attempts have been made to assess animal distress on the basis of behavioral or physiological responses rather than expert scoring. For example, Langford and colleagues (Langford et al., 2010) developed the mouse grimace scale for assessing degrees of pain from the facial expressions of the laboratory mouse. Changes in facial expression correlated in a dose dependent manner with other pain assessments, such as the pain reflex responses employed in preclinical pain research, and the painful face disappeared upon administration of morphine. Physiological measures utilized to assess the pain and distress associated with animal experiments include non-invasive measurement of stress hormone metabolites in feces and of amylase levels in saliva (Kolbe et al., 2015; Matsuura et al., 2012). However, background knowledge of the metabolism and excretion of metabolites in relationship to age, sex, strain and circadian rhythm is required, and each experimental situation must be carefully validated with respect to sampling and analytical procedures. In addition, although these approaches may provide more objective information about animal distress, this information cannot be added up in a meaningful way and thus cannot be used to directly estimate ethical costs.

When ethical weights are assigned cardinal properties, it becomes possible to give the experience suffered by each animal a numeric value. These weights could then be applied to different study designs in toxicity testing where defined dose groups are used and the risk of clinical signs of toxicity is dose dependent. Consider a traditional cancerogenicity test with equally sized dose groups, e.g., 4 dose groups with 50 animals per group (control-low-mid-high doses). In this case the probability of experiencing a relatively mild clinical sign with an ethical weight of 5 (corresponding to a 10% weight loss here) is 0%, 5%, 15% and 40% in the control, low, medium and high dose group, respectively. The presumed ethical weight for an individual animal can then be calculated as follows:

$$Ethical\ weight = p_{sign} \times w_{sign} + (1 - p_{sign}) \times w_{nosign}$$

where p_{sign} is the probability that an animal experiences a specific clinical sign, w_{sign} is the ethical weight of that sign and w_{nosign} is the ethical weight of an animal that does not experience any clinical signs. Accordingly, the weights will be 1 for a control animal, 1.2 for a low dose animal, 1.6 for a mid dose animal and 2.6 for a high dose animal, and the total estimated suffering would be 320 ($50 \times 1 + 50 \times 1.2 + 50 \times 1.6 + 50 \times 2.6$). With such an approach, ethical weights can provide an additional tool for optimal design of experiments. Instead of designing experiments to have as few animals as possible they will be designed to minimize the total ethical cost (Öberg, 2010).

Others have examined the cost-effectiveness of various test strategies with respect to regulatory (Nordberg et al., 2008) or economic (Gabbert and van Ierland, 2010) value. Nordberg and colleagues concluded that within the classification and labelling system, to perform short-term testing of a larger number of substances appears to be more efficient than to perform sub-acute toxicity studies on substances already tested for acute

toxicity. It remains to be determined how ethical weights might influence this conclusion. Gabbert and van Ierland discuss how analysis of cost-effectiveness can include both economic and animal welfare goals, and these investigators have developed a three-dimensional extension of the standard cost-effectiveness analysis. However, in their model the total number of animals served as a proxy for animal welfare.

The ethical weights derived in the current investigation focus on single clinical signs in toxicity testing. Future research may cover more symptoms as well as combinations of symptoms. A similar procedure for deriving ethical weights might also be applied to the vast majority of animal-based experiments, e.g., animal models of disease. In that case, the ethical weights can also be used in post-experimental evaluations, for instance to estimate the total ethical cost of animal experiments.

Quantitative scoring schemes have met resistance on the grounds that harms and benefits are not quantifiable and comparable in the necessary way (Vieira de Castro and Olsson, 2015). Quantitative scoring schemes have also been criticized for giving a false sense of accuracy and objectivity (Voipio et al., 2004). Our ethical weights can of course be put to question, but are less open for the criticism regarding comparability, relative to the scoring schemes proposed by others, as our weights do not compare harms and benefits. Our weights presented are however equally open for criticism concerning accuracy, and the variation in the answers given by the participants indicates that they are not very accurate. On the other hand, qualitative statements about reduction and refinement lack specificity, transparency and/or evaluability. Also, as we have shown with examples, qualitative statements are not sufficient to guide experimental design and alternative assessments.

Trade-off methodologies, such as the one employed here, do indeed have weaknesses, for instance they are prone to be dependent on numerical anchoring (Ubel et al., 2002). That is, answers can be anchored on earlier answers or numbers given by the interviewer. In the present study, participants could not anchor their answer to a number given by the interviewer but rather had to state their point-of-indifference directly. On the other hand, directly stating a point-of-indifference can also be difficult. Also, an obvious limitation with the ethical weights is that they are strictly descriptive and empirical. They are solely based on answers given by members of the ethical committees, and not directly supported by normative arguments.

5 Conclusion

In summary, ethical weights were derived for typical clinical signs of toxicity associated with testing involving laboratory rodents by asking members of Animal Ethics Committees in Sweden to grade signs and symptoms of distress. There were no significant differences between the weights assigned by political nominees, representatives of animal welfare organizations and researchers. Body weight loss was rated as the least severe outcome, and hunched posture was considered the most severe. We propose that ethical weights with cardinal properties

could be used as systematic and transparent tools to account for both the number of animals used and the amount of distress or suffering they are subjected to. Such weights would allow the inclusion of animal distress in cost-effectiveness analyses of experimental design and testing strategies and potentially also in both pre- and postexperimental ethical assessment of animal experimentation.

References

Chapman, K., Sewell, F., Allais, L. et al. (2013). A global pharmaceutical company initiative: An evidence-based approach to define the upper limit of body weight loss in short term toxicity studies. *Regul Toxicol Pharmacol 67*, 27-38. https://doi.org/10.1016/j.yrtph.2013.04.003

Damschroder, L. J., Roberts, T. R., Zikmund-Fisher, B. J. et al. (2007). Why people refuse to make tradeoffs in person tradeoff elicitations: A matter of perspective? *Med Decis Making 27*, 266-280. https://doi.org/10.1177/0272989X07300601

de Boo, M. J., Rennie, A. E., Buchanan-Smith, H. M. et al. (2005). The interplay between replacement, reduction and refinement: Considerations where the Three Rs interact. *Anim Welfare 14*, 327-332.

Delpire, V. C., Ben Mepham, T. and Balls, M. (1999). A proposal for a new ethical scheme addressing the use of laboratory animals for biomedical purposes. *Altern Lab Anim 27*, 869-881.

EC (2006). 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. http://bit.ly/2hiiOZL

EC (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. http://bit.ly/2gvJszu

EC (2010). Caring for animals aiming for better science. Directive 2010/63/eu on protection of animals used for scientific purposes. http://ec.europa.eu/environment/chemicals/lab_animals/pdf/guidance/inspections/en.pdf

EC (2013). Seventh Report from the Commission to the Council and the European Parliament on the Statistics on the number of animals used for experimental and other scientific purposes in the member states of the European Union. http://ec.europa.eu/environment/chemicals/lab_animals/reports_en.htm (accessed 25.06.2016)

EU (2010). Directive 2010/63/EU of the European Parliament and of the council of 22 September 2010 on the protection of animals used for scientific purposes. *Off J Eur Union L 276*, 33-79.

Franco, N. H. and 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. *Lab Anim 48*, 50-60. https://doi.org/10.1177/0023677213498717

Franco, N. H., Magalhães-Sant´Ana, M. and Olsson, A. S. (2014). Welfare and quantity of life. In M. C. Appleby, D. M. Weary and P. Sandøe (eds.), *Dilemmas in Animal Welfare*. London: CABI. https://doi.org/10.1079/9781780642161.0046

Gabbert, S. and van Ierland, E. C. (2010). Cost-effectiveness analysis of chemical testing for decision-support: How to include animal welfare? *Hum Ecol Risk Assess 16*, 603-620. https://doi.org/10.1080/10807031003788840

Gold, M. R., Stevenson, D. and Fryback, D. G. (2002). HALYs and QALYs and DALYs, oh my: Similarities and differences in summary measures of population health. *Annu Rev Public Health 23*, 115-134. https://doi.org/10.1146/annurev.publhealth.23.100901.140513

Hansson, S. O. (2014). The moral oracle's test. *Ethical Theory and Moral Practice 17*, 643-651. https://doi.org/10.1007/s10677-014-9500-5

Karolinska Institutet (2016). The KI Assessment Checklist. https://internwebben.ki.se/sites/default/files/ki_bedomningsmall_eng.pdf (accessed 29.03.2016)

Kolbe, T., Palme, R., Tichy, A. et al. (2015). Lifetime dependent variation of stress hormone metabolites in feces of two laboratory mouse strains. *PLoS One 10*, e0136112. https://doi.org/10.1371/journal.pone.0136112

Langford, D. J., Bailey, A. L., Chanda, M. L. et al. (2010). Coding of facial expressions of pain in the laboratory mouse. *Nat Methods 7*, 447-449. https://doi.org/10.1038/nmeth.1455

Leach, M. C., Thornton, P. D. and Main, D. C. J. (2008). Identification of appropriate measures for the assessment of laboratory mouse welfare. *Anim Welfare 17*, 161-170.

Matsuura, T., Takimura, R., Yamaguchi, M. et al. (2012). Estimation of restraint stress in rats using salivary amylase activity. *J Physiol Sci 62*, 421-427. https://doi.org/10.1007/s12576-012-0219-6

Morton, D. B. and Griffiths, P. H. M. (1985). Guidelines on the recognition of pain, distress and discomfort in experimental-animals and an hypothesis for assessment. *Vet Rec 116*, 431-436. https://doi.org/10.1136/vr.116.16.431

NC3R/LASA (2009). Guidance on dose level selection for regulatory general toxicology studies for pharmaceuticals. http://www.lasa.co.uk/pdf/lasa-nc3rsdoselevelselection.pdf

Nord, E. (1995). The person-trade-off approach to valuing health-care programs. *Med Decis Making 15*, 201-208. https://doi.org/10.1177/0272989X9501500302

Nordberg, A., Ruden, C. and Hansson, S. O. (2008). Towards more efficient testing strategies – analyzing the efficiency of toxicity data requirements in relation to the criteria for classification and labelling. *Regul Toxicol Pharmacol 50*, 412-419. https://doi.org/10.1016/j.yrtph.2008.01.012

Öberg, M. (2010). Benchmark dose approaches in chemical health risk assessment in relation to number and distress of laboratory animals. *Regul Toxicol Pharmacol 58*, 451-454. https://doi.org/10.1016/j.yrtph.2010.08.015

Porter, D. G. (1992). Ethical scores for animal-experiments. *Nature 356*, 101-102. https://doi.org/10.1038/356101a0

Prades, J. L. P. (1997). Is the person trade-off a valid method for allocating health care resources? *Health Econ 6*, 71-81. https://doi.org/10.1002/(SICI)1099-1050(199701)6:1<71::AID-HEC239>3.0.CO;2-Z

Purves, K. E. (2000). Global harmonisation of pain and distress classification systems: Current analysis and opportunity for change. *Progress in the Reduction, Refinement and Replacement of Animal Experimentation 31*, 1057-1067.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org

Rawls, J. A. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

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

SAMS (2005). *Ethical Principles and Guidelines for Experiments on Animals*. 3rd edition.

Sandoe, P. and Christiansen, S. B. (2007). The value of animal life: How should we balance quality against quantity? *Anim Welfare 16*, 109-115.

Scharmann, W. (1999). Physiological and ethological aspects of assessment of pain, distress and suffering. In C. F. M. Hendriksen and D. B. Morton (eds.), *In Humane Endpoints in Animal Experiments for Biomedical Research*. London: Royal Society of Medicine Press.

Shao, K. and Small, M. J. (2012). Statistical evaluation of toxicological experimental design for Bayesian model averaged benchmark dose estimation with dichotomous data. *Hum Ecol Risk Assess 18*, 1096-1119. https://doi.org/10.1080/10807039.2012.707938

Slob, W. (2014). Benchmark dose and the three Rs. Part I. Getting more information from the same number of animals. *Crit Rev Toxicol 44*, 557-567. https://doi.org/10.3109/10408444.2014.925423

Stafleu, F. R., Tramper, R., Vorstenbosch, J. et al. (1999). The ethical acceptability of animal experiments: A proposal for a system to support decision-making. *Lab Anim 33*, 295-303. https://doi.org/10.1258/002367799780578255

Ubel, P. A., Richardson, J. and Baron, J. (2002). Exploring the role of order effects in person trade-off elicitations. *Health Policy 61*, 189-199. https://doi.org/10.1016/S0168-8510(01)00238-X

Vieira de Castro, A. C. and Olsson, A. S. (2015). Does the goal justify the methods? Harm and benefit in neuroscience research using animals. In G. Lee, J. Illes and F. Ohl (eds.), *Ethical Issues in Behavioral Neuroscience*. Berlin: Springer.

Voipio, H.-M., Kaliste, E., Hirsjärvi, P. et al. (2004). Nordic-European workshop on ethical evaluation of animal experiments. *Scand J Lab Anim Sci 31*, 251-267.

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

Correspondence to
Joakim Ringblom, PhD
Karolinska Institutet
Institute of Environmental Medicine
P.O. Box 210, SE-171 77
Stockholm, Sweden
Phone: +46 8 524 875 48
e-mail: joakim.ringblom@ki.se